TURING

Prentice Hall



运筹学导论

高级篇

运筹学导论：高级篇（第8版）

Operations Research: An Introduction

“本书全面地论述了运筹学的三个方面——理论、应用和计算，而且游刃有余。我求学时就是通过本书老版本学习运筹学的，如今我在使用新版本教授学生，这么多年了，它仍然是本领域的经典。”

——Amazon.com

运筹学是一门应用领域十分广泛的学科，它应用分析、试验、量化的方法，对经济管理系统中人力、物力、财力等资源进行统筹安排，为决策者提供有依据的最佳方案，以实现最有效的管理。

本书是运筹学方面的经典著作之一，理论严密，案例丰富，并且充分运用了计算机软件，体现了作者在运筹学教学研究和业界实践方面精湛的造诣，已被翻译成中、韩、西班牙、日、俄、土耳其、印尼、马来等多种语言，为全球众多高校采用，深受好评。第8版对教材内容作了较大的修订，在教材的编排上突出反映运筹学中的应用问题和计算方法。

本书特色

- 理论联系实际，应用色彩浓厚。
- 注重与计算机软件程序相结合，富有时代气息。
- 每章开头都有本章导读，帮助读者了解教材内容。

将原书分成两册出版后，对原书章节顺序进行了调整。初级篇内容全面，符合国内的大纲要求，可作为相关专业本科生教材。高级篇可供研究生、MBA作为教材或者参考书。



Hamdy A. Taha 美国阿肯色大学荣休教授，世界知名运筹学家。曾在全球各地任教和担任顾问，拥有非常丰富的教学研究和实践经验。他在*Management Science*和*Operations Research*等世界顶级学术刊物上发表了大量论文。

* 封面图片为清代画家苏六朋创作的《东山报捷图》，作品描绘的是东晋名士谢安处理淝水大战之际，镇静自若与友弈棋的情形。《资治通鉴》如此描述：谢安得驿书，知秦兵已败，时方与客围棋，摄书置床上，了无喜色，围棋如故。客问之，徐答曰：“小儿辈遂已破贼。”既罢，还内，过户限，不觉屐齿之折。





www.PearsonEd.com

本书相关信息请访问：图灵网站 http://

读者热线：(010)88593802

反馈/投稿/推荐信箱：contact@turingbook.com

分类建议 数学/运筹学

人民邮电出版社网址 www.ptpress.com.cn

ISBN 978-7-115-18947-9

9 787115 189479 >

ISBN 978-7-115-18947-9/O1

定价：59.00 元

TURING



运筹学导论

高级篇

图书在版编目(CIP)数据

运筹学导论：第8版，高级篇/(美)塔哈(Taha, H. A.)著；薛毅等译. —北京：人民邮电出版社，2008. 12
(图灵数学·统计学丛书)
书名原文：Operations Research: An Introduction
ISBN 978-7-115-18947-9

Ⅰ. 运… Ⅱ. ①塔… ②薛… Ⅲ. 运筹学 Ⅳ. O22

中国版本图书馆CIP数据核字(2008)第154117号

内 容 提 要

本书是运筹学方面的经典著作之一，为全球众多高校采用. 高级篇共12章，内容包括高级线性规划、概率论基础复习、随机库存模型、仿真模型、马尔可夫链、经典最优化理论、非线性规划算法、网络和线性规划算法进阶、预测模型、随机动态规划、马尔可夫决策过程、案例分析等，并附有统计表、部分习题答案、向量和矩阵复习，以及应用案例.

本书可作为高等院校经管类专业和数学专业的教材，也可供MBA及相关研究人员参考.

图灵数学·统计学丛书

运筹学导论：高级篇(第8版)

◆ 著　[美] Hamdy A. Taha
译　薛　毅　刘德刚　朱建明　侯思祥
审校　韩继业
责任编辑　明永玲
执行编辑　张继发

◆ 人民邮电出版社出版发行　北京市崇文区夕照寺街14号
邮编　100061　电子函件　315@ptpress.com.cn
网址：http://www.ptpress.com.cn
北京铭成印刷有限公司印刷

◆ 开本：700×1000　1/16
印张：24.5
字数：521千字　2008年12月第1版
印数：1–3 000册　2008年12月北京第1次印刷

著作权合同登记号　图字：01-2006-5778号

ISBN 978-7-115-18947-9/O1

定价：59.00元

读者服务热线：(010)88593802　印装质量热线：(010) 67129223

反盗版热线：(010)67171154

版权声明

译 者 序

运筹学起源于 20 世纪二次大战期间, 是一门应用性很强的学科. 1938 年, 英国皇家空军部门在 Bawdsey 成立了一个从事作战研究的科学家小组, 小组的科学家把他们的研究工作称为 “operational research” (“operation” 在军事术语中意为 “作战”). 这是 “运筹学” 一词最早出现于文献的时间. “二战” 中英军每一个大的指挥部大都成立了这种运筹研究小组. 之后, 美国和加拿大的军事部门也成立了若干运筹研究小组 (美国称这种研究工作为 “operations research”). 他们广泛地研究有关战果评价、战术革新、技术援助、战略选择和战术计划等问题. “二战” 期间英、美、加等国军事部门的运筹研究小组的工作为同盟国战胜德、意、日等轴心国做出了卓越的贡献. 对于人类社会的科学进程而言, 这些科学家的集体工作和智慧开创了一门崭新的学科 —— 运筹学.

体现运筹学思想和方法的某些早期先驱性的研究工作, 可以追溯到 20 世纪初期. 例如, 1908 年丹麦工程师埃尔朗提出的电话话务理论是运筹学中排队论 (queueing theory) 的起源; 1916 年英国的兰彻斯特提出的战斗模型方程是军事运筹学早期的一项重要成果; 1939 年前苏联数学家坎托罗维奇在 *The Mathematical Method of Production Planning and Organization* 一书中, 开创性地提出线性规划, 并研究了工业生产的资源合理利用和计划等问题, 这一卓越贡献使他获得了 1975 年诺贝尔经济学奖; 基本的对策均衡的思想可追溯到 1838 年库尔诺的文章, 1913 年的德国策梅洛提出了抽象战略对策的数学模型, 1928 年冯 · 诺依曼提出了二人零和对策的解的一般理论, 这些是关于对策论的早期的研究. 上述这些先驱性成就对以后运筹学的发展有着深远的影响.

“二战” 以后, 美国等国家的军事部门保留和调整了运筹研究组, 人员编制得到了扩大, 运筹学有了新的发展. 1949 年美国成立了著名的兰德 (RAND) 公司. 与此同时, 许多运筹学工作者从军方转入企业、大学或政府部门. 在新的更宽阔的环境中, 运筹学的应用研究和理论研究得到了蓬勃发展, 多年来它已为欧美等国创造了数以亿计的社会财富.

简略地说, 运筹学的研究对象是现实世界中的运行系统, 这些运行系统的设计和运转受到管理人员的决策的影响和作用. 运筹学创造出一些理论 (包括数学模型) 和方法, 用来描述与分析这些运行系统的现象、性质和变化, 以寻求影响和作用于运行系统的设计与运转的最有效 (最优) 的决策, 发挥有限资源的最大效益, 使得运行系统达到总体最优的目标.

半个世纪以来, 运筹学在研究与解决各种复杂的实际问题中不断地得到创新和发展, 新模型、新理论和新方法不断涌现, 至今它已成为一个庞大的学科, 包括线性的和非线性的、连续的和离散的、确定性的和不确定性的许多分支. 运筹学的基本方法中有数学方法、统计学方法、仿真 (模拟) 方法、计算机科学方法等, 其中各种优化方法处于非常重要的地位.

运筹学的非凡价值, 使得在许多国家的大学里的运筹学系、管理科学系、经济学系、工业工程系、系统科学系、数学系、计算科学系等早已开设了关于运筹学及其一些分支学科的课程. 我国的情况也大致如此. 从适应于不同院系专业的学生学习运筹学来考虑, 一本好的运筹学基础的教科书十分重要. 在国外关于运筹学基础的诸多教材中, Hamdy A. Taha 著的 *Operations Research, An Introduction* 是非常优秀的一本. Taha 是美国阿肯色大学工业学院工业工程教授, 世界知名运筹学家. 他著的这本书最早出版于 1968 年. 多年来此书经过多次修改与扩增, 今年已出版了第 8 版. 本书被世界上多所大学采用为运筹学基础教材, 已有西班牙文、日文、俄文、土耳其文、印尼文等多种译本出版. 它有三大重要特色. 一是内容广泛, 取材得当. 连同附带的光盘, 共有 24 章和 5 个附录, 内容涉及线性规划、运输问题、网络问题、目标规划、整数规划、动态规划、库存问题、非线性规划等确定性运筹分支, 以及随机动态规划、随机库存问题、排队系统、马尔可夫决策过程、决策分析、对策论、模拟问题、预报问题等随机性运筹分支. 这些内容覆盖了迄今运筹学所研究的大部分重要问题. 该书在取材上首先重视对上述运筹问题的基本知识的讲解, 但对某些问题也包括了较高深的内容, 以满足不同读者的需要. 二是突出实用性. 书中各章总是通过若干实际问题的求解来引导出所要讲的运筹问题的数学模型. 这既凸显出这些运筹问题的实际背景, 也可使读者学到如何进行建模. 第 24 章详细地介绍了 15 个实际应用案例, 运用了多种运筹学技术进行建模、数据采集以及求解计算等. 附录 E 中还收录了近 50 个应用例子. 作者精心收集和分析的这些实例来源于工业、商业、金融、社会、体育、娱乐等许多行业, 是很好的运筹学教学资料. 三是计算方法与软件相结合. 全书使用教学辅助软件 TORA、电子表格程序 Excel 及 AMPL 等. 读者可以利用这些软件工具对所学的模型和计算方法进行计算和检验.

在我国运筹学基础一类的图书拥有大量读者, 这类图书的累计销售量有的已达几十万册. 但国内目前这类书籍只有很少几种. 2006 年人民邮电出版社图灵公司邀我们翻译新出版的《运筹学导论 (第 8 版)》, 这体现了出版社对发展我国运筹学的重视. 由于书的篇幅宏大, 有 1 000 多页, 中译本分成上下两册出版, 同时将附录 C 和索引拆分到了上下两册, 其中附录 C 放在了图灵网站 (www.turingbook.com) 上供读者免费注册下载. 上册主要包括原书中属于基础部分的 12 章, 以及附录 A; 下册主要包括属于提高部分的 12 章, 以及附录 B, D, E. 每册均可用做一个学期的教材. 本书第 2, 3, 4, 13 章及附录 A 由薛毅教授翻译, 第 5 章由侯思祥教授翻译, 第 6, 7, 8, 16, 20 章及附录 C 由朱建明博士翻译, 其余各章及附录 B, D, E 由刘德刚博士翻译, 全部译稿由我校阅. 中译本难免有疏漏和翻译不妥之处, 敬请读者给予指正.

韩继业

2007 年 5 月于中科院

前　言

本书第 8 版对教材内容作了很多的修订, 在教材的编排上突出反映运筹学中的应用问题和计算方法.

- 第 2 章通过城市规划、货币套利交易、投资、生产计划、混合配比、排序以及下料等实际问题的应用, 主要介绍了线性规划的建模. 新增加的节后习题也涉及从水质管理、交通控制到军事领域等多个运筹问题.
- 第 3 章以一种简单和直接的方式介绍了一般性的线性规划灵敏度分析, 包括对偶价格和简约费用, 作为单纯性表计算部分的直接扩充.
- 本版的第 4 章主要是基于对偶性进行线性规划后最优分析.
- 针对旅行商问题 (Traveling Salesperson Problem, TSP), 介绍了一个基于 Excel 的组合式最近邻点反向启发式算法.
- 新增的第 17 章扩充了马尔可夫链的处理方法.
- 在全新的第 24 章里, 详细介绍了 15 个实际应用案例. 对这些案例的分析通常涉及多种 OR 技术 (例如启发式算法和线性规划, 或者整数线性规划和排队论), 用来进行建模、数据采集以及问题的求解计算等. 这些应用问题在相关的各章里都有引用, 让读者能够充分了解在实际生活中如何运用运筹学技术.
- 新的附录 E 收录了按照章节排列的约 50 个小型实用问题的例子.
- 本书还包含了 1 000 多个节后习题, 其中题前标有星号 (*) 的表示附录 C 给出了相应的答案.
- 每章开头都有**本章导读**, 帮助读者了解教材内容, 有效利用附带的软件程序.
- 把教材与软件相结合可以让读者对需要深入介绍的概念进行实际检验.
  1. 全书都用到了 Excel 程序, 包括动态规划、旅行商问题、库存问题、层次分析法、贝叶斯概率、“电子化” 统计表、排队问题、模拟、马尔可夫链以及非线性规划等. 一些程序中的交互式用户输入功能有助于对相应方法的更好理解.
  2. 对 Excel 规划求解程序的使用扩展到了全书, 特别用在线性规划、网络规划、整数规划和非线性规划问题.
  3. AMPL® 是一种强大的商业化建模语言, 本书将 AMPL 结合在大量的例题中, 这些例子涉及线性、网络、整数和非线性规划问题. 附录 A 给出了 AMPL 的语句规则以及本书例题中所引用的语言素材.
  4. 本书中, TORA 仍然充当教学软件的重要角色.
- 所有与计算机相关的材料都相对独立, 有的作为单独的章节, 有的按照标题 AMPL/Excel/Solver/TORA 程序作为一小节, 以尽量不影响本书的主要内容的介绍.

为了限制本书的页数, 我们把一些小节、一部分整章以及两个附录都放在了附带的光盘里. 作者根据运筹学导论课程中不太经常用到的内容截选下来, 放在光盘里.[①]

致 谢

我首先要感谢新泽西理工学院的 Layek L. Abdel-Malek、路易斯安那州立大学的 Evangelos Triantaphyllou、俄克拉荷马州立大学的 Michael Branson、中央佛罗里达大学的 Charles H. Reilly 以及弗吉尼亚技术学院及州立大学的 Mazen Arafeh, 他们对本书的第 7 版提供了重要的修改意见. 我要特别感谢下面两位学者, 因为他们让我在第 8 版写作期间的思路受到了很大启迪: R. Michael Harnett (堪萨斯州立大学) 多年来在本书的组织和内容方面给我提供了许多宝贵的建议; Richard H. Bernhard (北卡罗来纳州立大学) 对第 7 版所提出的详细改进意见促使本版对前面若干章做了重新调整.

Robert Fourer (西北大学) 针对本版介绍的 AMPL 内容耐心地提出了宝贵的建议, 我衷心地感谢他对这些内容的编辑所提供的帮助, 并感谢他的修改建议, 使得这部分内容读起来更通畅. 我还要感谢他的帮忙, 让我们能在附带的光盘中包含 AMPL 学生版以及 CPLEX、KNITRO、LPSOLVE、LOQO 以及 MINOS 等求解程序.[②]

我一直对我的同事们以及很多的学生们深表感激, 感谢他们的建议和鼓励. 我这里要特别感谢 Yuh-Wen Chen (台湾大叶大学)、Miguel Crispin (得克萨斯大学 El Paso 分校)、David Elizandro (田纳西技术大学)、Rafael Gutiérrez (得克萨斯大学 El Paso 分校)、Yasser Hosni (中央佛罗里达大学)、Erhan Kutanoglu (得克萨斯大学奥斯汀分校)、Robert E. Lewis (美军管理工程学院)、Gino Lim (休斯顿大学)、Scott Mason (阿肯色大学)、Heather Nachtman (阿肯色大学)、Manuel Rossetti (阿肯色大学)、Tarek Taha (JB Hunt 公司), 以及 Nabeel Yousef (中央佛罗里达大学), 谢谢他们对我的支持和帮助.

我还要对 Pearson Prentice Hall 出版社的编辑出版组的 Dee Bernhard (副总编)、David George (生产经理)、Bob Lentz (文字编辑)、Craig Little (技术编辑), 以及 Holly Stark (高级组稿编辑) 表达我的衷心谢意, 感谢他们在本书出版期间的出色工作.

HAMDY A. TAHA
hat@uark.edu
http://ineg.uark.edu/TahaORbook/

① 光盘中的这些内容已放在本书的高级篇中. —— 编者注
② 这些内容可到图灵网站 www.turingbook.com 下载. —— 编者注

目　录

① 附录 A 在本书的初级篇中. —— 编者注
② 附录 C(上) 属本书初级篇. 读者均可到图灵网站免费注册下载. —— 编者注

(初级篇) 目 录

第 13 章　高级线性规划

本章导读　本章介绍线性规划与对偶理论的数学基础，讲述许多简洁高效的算法，包括修正单纯形法、有界变量和参数规划. 第 20 章将介绍处理大规模线性规划的两种附加算法：分解算法和 Karmarkar 内点算法.

本章的内容用到了很多矩阵代数的知识. 附录 D 复习了矩阵代数的内容.

本章应该特别注意的 3 个要点是：修正单纯形法、有界变量算法和参数规划. 在修正单纯形法中，矩阵代数用来很好地控制机器舍入误差，这个问题在第 3 章介绍的行运算方法中曾经提到过. 有界变量算法在整数规划的分支定界算法 (见第 8 章) 中起着重要的作用. 参数规划在线性规划模型上添加了动态参数，从而能够得出模型的参数连续变化后所带来的最优解的变化结果.

总结第 3 章中的以简单易懂的单纯形表形式所表示的矩阵运算结果，将有助于深入理解修正单纯形算法、有界变量算法、分解算法以及参数规划的细节内容. 尽管采用矩阵运算使算法显得似乎不同，但其理论与第 3 章的内容完全一致.

本章共包括 1 个实际问题的应用、8 个例题、58 个节后习题和 4 个章后综合问题. 综合问题汇总在附录 E 中. AMPL/Excel/Solver/TORA 程序在文件夹 ch13Files 中.

实际应用——泰国海军运送新兵最优行船路线与人员指派

泰国海军每年征兵 4 次. 招募的新兵到全国 34 个征兵中心报到，然后由汽车运送到 4 个海军分基地. 从那里，再用船将新兵们送到海军总基地. 分基地的码头对靠岸的船型有一定的限制. 分基地的运送能力是有限的，但就整体而言，4 个分基地有充分的能力运送所有的应征者. 在 1983 年夏季，共有 2 929 名新兵从征兵中心被送到 4 个分基地，最终被送到总基地. 问题是要确定运送新兵的最优计划安排，首先从征兵中心到分基地，然后从分基地到总基地. 这项研究综合运用了线性规划与整数规划. 其具体细节见第 24 章的案例 5.

13.1　单纯形法的基本原理

在线性规划中，如果连接任意两个不同可行点所形成的线段还落在可行集合内，则称可行解空间构成一个**凸集**(convex set). **凸集的极点** (extreme point) 是一个可行点，但不能位于连接可行集合中任意两个不同可行点的线段上. 实际上，极

点与角点是相同的, 但角点用于第 2 章至第 4 章更为合适.

图 13.1 画出了两个集合. 集合 (a), 线性规划解空间的典型形式, 是一个 (具有 6 个极点的) 凸集, 而集合 (b) 不是凸集.

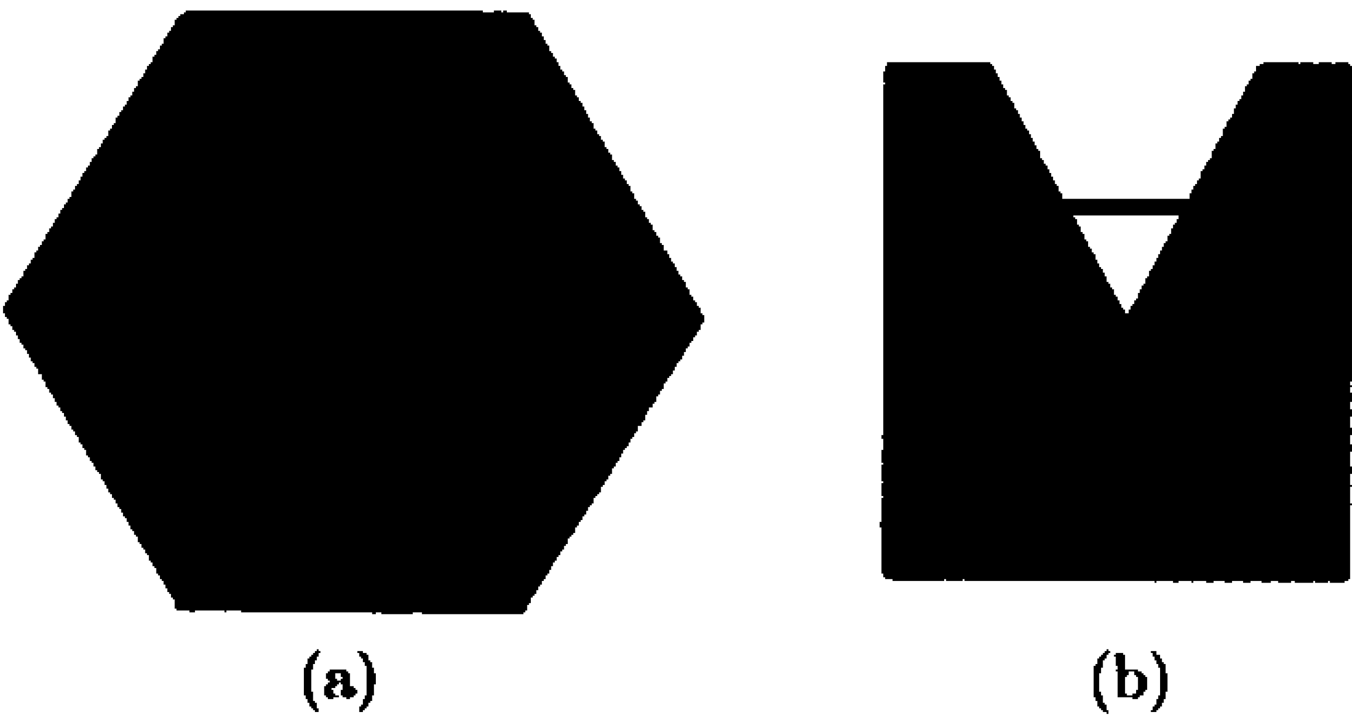


图 13.1　凸集与非凸集的例子

在 2.3 节介绍的线性规划的图解法中, 我们说明了, 最优解总是与解空间中某个可行的极点 (角点) 相关. 这个结果有着直观的意义, 因为在线性规划解空间中, 每个可行点能够由其可行极点的函数来确定. 例如, 在图 13.1 的凸集 (a) 中, 可行点 X 可以表示成其极点 X_1, X_2, X_3, X_4, X_5, X_6 的**凸组合** (convex combination), 其表达式为

$$X = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6$$

其中

$$\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 + \alpha_6 = 1$$

$$\alpha_i \geqslant 0, \quad i = 1, 2, \cdots, 6$$

这个观察结果说明, 解空间是完全可以由极点来确定的.

例 13.1-1

证明: 集合

$$C = \{(x_1, x_2) \mid x_1 \leqslant 2, x_2 \leqslant 3, x_1 \geqslant 0, x_2 \geqslant 0\}$$

是凸集.

令 $X_1 = (x_1', x_2')$ 和 $X_2 = (x_1'', x_2'')$ 是 C 中任意两个不同的点. 如果 C 是凸集, 则 $X = (x_1, x_2) = \alpha_1 X_1 + \alpha_2 X_2$, $\alpha_1 + \alpha_2 = 1$, $\alpha_1, \alpha_2 \geqslant 0$ 必属于 C. 为了证明这是对的, 需要证明线段 X 满足 C 的所有约束, 即

$$x_1 = \alpha_1 x_1' + \alpha_2 x_1'' \leqslant \alpha_1(2) + \alpha_2(2) = 2$$

$$x_2 = \alpha_1 x_2' + \alpha_2 x_2'' \leqslant \alpha_1(3) + \alpha_2(3) = 3$$

因此, $x_1 \leqslant 2$ 且 $x_2 \leqslant 3$. 此外, 非负条件满足, 因为 α_1 和 α_2 都是非负的.

习题 13.1A

1. 证明: 集合 $Q = \{(x_1, x_2) \mid x_1 + x_2 \leqslant 1, x_1 \geqslant 0, x_2 \geqslant 0\}$ 是凸集. 非负条件是证明所必需的吗?

*2. 证明：集合 $Q=\{(x_1,x_2)\mid x_1\geqslant 1$ 或 $x_2\geqslant 2\}$ 不是凸集.

3. 用图形确定下面凸集的极点：

$$Q=\{(x_1,x_2)\mid x_1+x_2\leqslant 2, x_1\geqslant 0, x_2\geqslant 0\}$$

证明：整个解空间能够由其极点的凸组合来确定. 因此可知, 一旦解空间的极点已知, 任意凸的 (有界) 解空间就被完全确定.

4. 在图 13.2(画有刻度) 的解空间中, 将内点 (3, 1) 表示成极点 A, B, C, D 的凸组合, 而且每个极点的权数都严格为正.

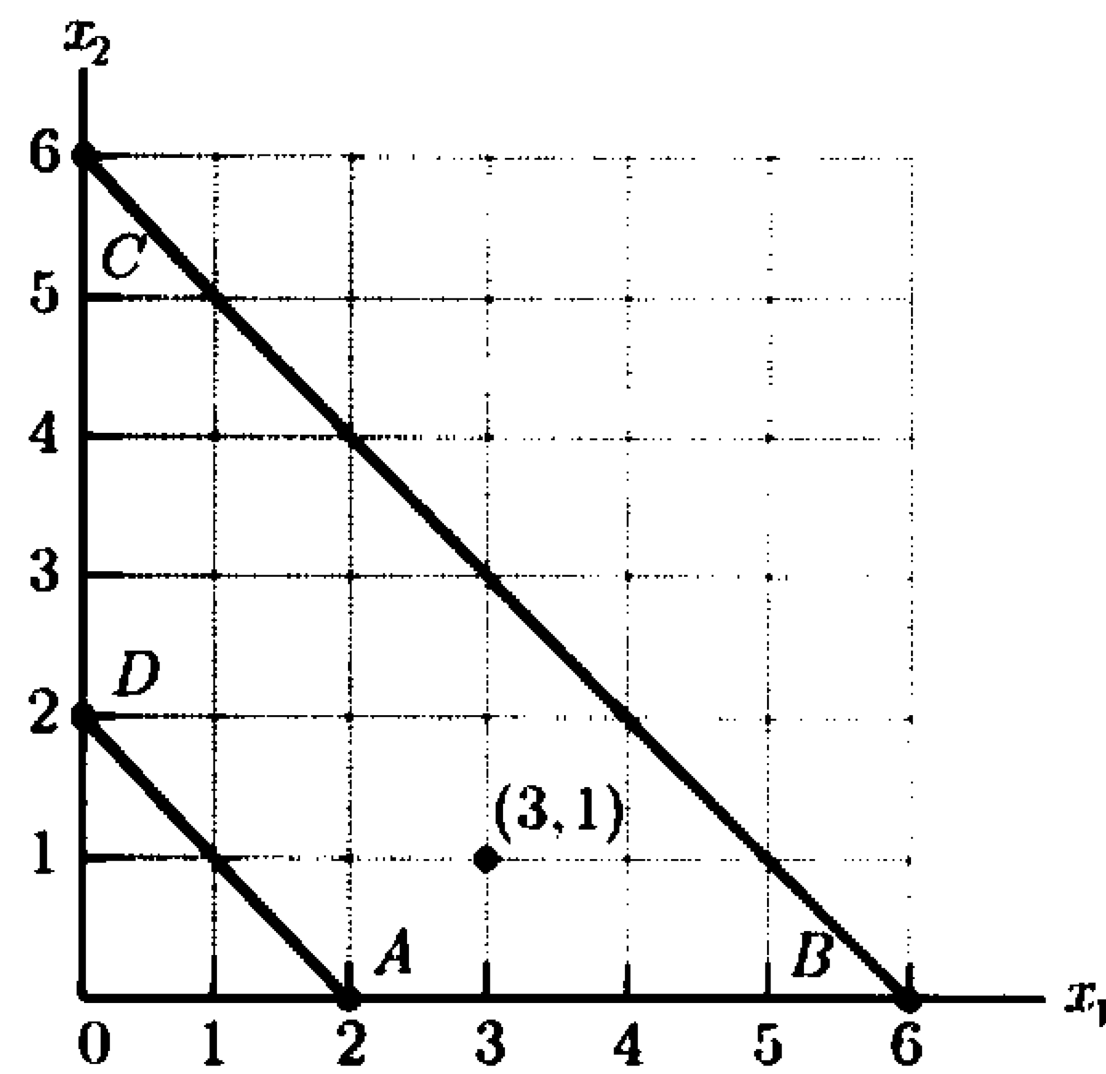


图 13.2 习题 13.1A 第 4 题的解空间

13.1.1 从极点到基本解

用矩阵记号表达具有等式形式的一般线性规划问题 (见 3.1 节) 是很方便的. 定义 $\boldsymbol{X}$ 为 n 维向量, 表示问题的变量; $\boldsymbol{A}$ 是 $m\times n$ 阶矩阵, 表示约束的系数; $\boldsymbol{b}$ 为列向量, 表示约束的右端项; $\boldsymbol{C}$ 是 n 维向量, 表示目标函数系数. 则线性规划可写成

$$\begin{aligned} \max\text{或}\min \quad & z=\boldsymbol{CX} \\ \text{s.t.} \quad & \boldsymbol{AX}=\boldsymbol{b} \\ & \boldsymbol{X}\geqslant \boldsymbol{0} \end{aligned}$$

用第 3 章的格式 (见图 4.1), 通过适当排列初始基本解中相应的松弛变量或人工变量, 总可以使矩阵 $\boldsymbol{A}$ 的最右边 m 列构成单位矩阵 $\boldsymbol{I}$.

$\boldsymbol{AX}=\boldsymbol{b}$ 的**基本解** (basic solution) 的定义如下：令方程中的 $n-m$ 个变量等于零, 然后求解其余具有 m 个未知量的 m 个方程, **倘若得到的解是唯一的**, 则其解为基本解. 有了这个定义, 线性规划的理论建立了极点的几何定义与基本解的代数定义之间的关系：

$$\{\boldsymbol{X}|\boldsymbol{AX}=\boldsymbol{b}\}\text{的极点} \Leftrightarrow \boldsymbol{AX}=\boldsymbol{b}\ \text{的基本解}$$

这个关系蕴涵着线性规划解空间的极点完全由方程组 $\boldsymbol{AX}=\boldsymbol{b}$ 的基本解确定, 反之也是如此. 因此, 可得如下结论: $\boldsymbol{AX}=\boldsymbol{b}$ 的基本解包含了我们需要确定的线性规划问题最优解的所有信息. 此外, 如果我们强制非负限制 $\boldsymbol{X}\geqslant \boldsymbol{0}$, 则最优解的搜索只限制在可行的基本解上.

为了用数学公式定义基本解, 将方程组 $\boldsymbol{AX}=\boldsymbol{b}$ 用向量的形式表示:

$$\sum_{j=1}^{n}\boldsymbol{P}_j x_j=\boldsymbol{b}$$

向量 $\boldsymbol{P}_j$ 是矩阵 $\boldsymbol{A}$ 的第 j 列. 称 m 个向量的子集构成一个基 (basis) $\boldsymbol{B}$, 当且仅当所选择的 m 个向量线性无关 (linearly independent). 在这种情况下, 矩阵 $\boldsymbol{B}$ 非奇异 (nonsingular). 如果 $\boldsymbol{X}_B$ 由 m 个变量构成, 其变量与非奇异矩阵 $\boldsymbol{B}$ 中的列向量相对应, 则称 $\boldsymbol{X}_B$ 为基本解. 在这种情况下, 我们有

$$\boldsymbol{B}\boldsymbol{X}_B=\boldsymbol{b}$$

已知 $\boldsymbol{B}$ 的逆 $\boldsymbol{B}^{-1}$, 则可得到相应的基本解为

$$\boldsymbol{X}_B=\boldsymbol{B}^{-1}\boldsymbol{b}$$

如果 $\boldsymbol{B}^{-1}\boldsymbol{b}\geqslant \boldsymbol{0}$, 则 $\boldsymbol{X}_B$ 是可行的. 规定余下的 $n-m$ 个变量是非基的 (nonbasic) 且取 0 值.

前面的结果表明, 在一个含有 m 个方程 n 个未知量的方程组中, (可行和不可行) 基本解的最大个数为

$$\binom{n}{m}=\frac{n!}{m!(n-m)!}$$

例 13.1-2

确定下面方程的所有基本解, 并分类 (可行和不可行).

$$\begin{pmatrix}1 & 3 & -1\\ 2 & -2 & -2\end{pmatrix}\begin{pmatrix}x_1\\ x_2\\ x_3\end{pmatrix}=\begin{pmatrix}4\\ 2\end{pmatrix}$$

下表概括了其结果. 使用附录 D.2.7 的方法之一确定 $\boldsymbol{B}$ 的逆.

$\boldsymbol{B}$	$\boldsymbol{B}\boldsymbol{X}_B=\boldsymbol{b}$	解	状态
$(\boldsymbol{P}_1,\boldsymbol{P}_2)$	$\begin{pmatrix}1 & 3\\ 2 & -2\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}4\\ 2\end{pmatrix}$	$\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}\frac{1}{4} & \frac{3}{8}\\ \frac{1}{4} & -\frac{1}{8}\end{pmatrix}\begin{pmatrix}4\\ 2\end{pmatrix}=\begin{pmatrix}\frac{7}{4}\\ \frac{3}{4}\end{pmatrix}$	可行
$(\boldsymbol{P}_1,\boldsymbol{P}_3)$	(基不存在)	—	—
$(\boldsymbol{P}_2,\boldsymbol{P}_3)$	$\begin{pmatrix}3 & -1\\ -2 & -2\end{pmatrix}\begin{pmatrix}x_2\\ x_3\end{pmatrix}=\begin{pmatrix}4\\ 2\end{pmatrix}$	$\begin{pmatrix}x_2\\ x_3\end{pmatrix}=\begin{pmatrix}\frac{1}{4} & -\frac{1}{8}\\ -\frac{1}{4} & -\frac{3}{8}\end{pmatrix}\begin{pmatrix}4\\ 2\end{pmatrix}=\begin{pmatrix}\frac{3}{4}\\ -\frac{7}{4}\end{pmatrix}$	不可行

还可以用下面向量形式的表达式来研究这个问题:

$$\begin{pmatrix}1\\2\end{pmatrix}x_1+\begin{pmatrix}3\\-2\end{pmatrix}x_2+\begin{pmatrix}-1\\-2\end{pmatrix}x_3=\begin{pmatrix}4\\2\end{pmatrix}$$

其中 $\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{P}_3$ 和 $\boldsymbol{b}$ 均为二维向量, 它们可以用一般形式 $(a_1,a_2)^{\mathrm{T}}$ 来表示. 图 13.3 给出了这些向量在 (a_1,a_2) 平面上的图形表示. 例如, 对于 $\boldsymbol{b}=(4,2)^{\mathrm{T}}$, 有 $a_1=4$ 和 $a_2=2$.

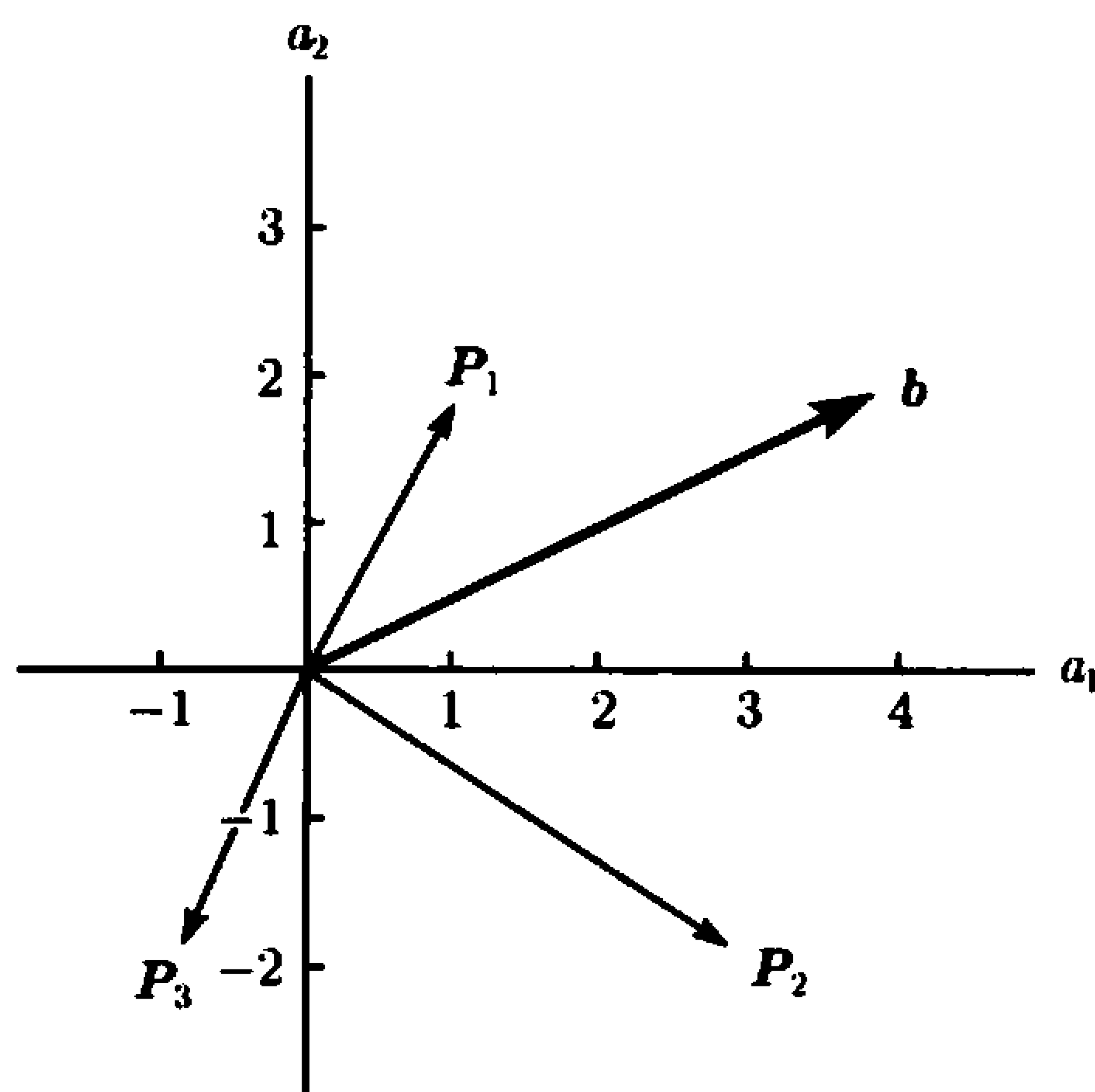


图 13.3 线性规划解空间的向量表示

因为我们处理的是一个二元 $(m=2)$ 方程, 所以一个基必须严格地包含两个向量, 且它们是从 $\boldsymbol{P}_1$, $\boldsymbol{P}_2$, $\boldsymbol{P}_3$ 中选出的. 从图 13.3 看到, 矩阵 $(\boldsymbol{P}_1,\boldsymbol{P}_2)$ 和 $(\boldsymbol{P}_2,\boldsymbol{P}_3)$ 构成基, 因为它们相对应的向量线性无关. 而矩阵 $(\boldsymbol{P}_1,\boldsymbol{P}_3)$ 对应的两个向量是相关的, 因此不能组成一个基.

从代数的观点来看, 一个 (方形) 矩阵, 如果它的行列式不为零, 则构成一个基 (见附录 D.2.5). 下面的计算说明组合 $(\boldsymbol{P}_1,\boldsymbol{P}_2)$ 和 $(\boldsymbol{P}_2,\boldsymbol{P}_3)$ 是基, 而组合 $(\boldsymbol{P}_1,\boldsymbol{P}_3)$ 不是基.

$$\det(\boldsymbol{P}_1,\boldsymbol{P}_2)=\det\begin{pmatrix}1&3\\2&-2\end{pmatrix}=1\times(-2)-2\times3=-8\neq0$$

$$\det(\boldsymbol{P}_2,\boldsymbol{P}_3)=\det\begin{pmatrix}3&-1\\-2&-2\end{pmatrix}=3\times(-2)-(-2)\times(-1)=-8\neq0$$

$$\det(\boldsymbol{P}_1,\boldsymbol{P}_3)=\det\begin{pmatrix}1&-1\\2&-2\end{pmatrix}=1\times(-2)-2\times(-1)=0$$

可以利用该问题的向量表示形式来讨论如何确定单纯形表的初始解. 由图 13.3 中的向量表示可知, 基 $\boldsymbol{B}=(\boldsymbol{P}_1,\boldsymbol{P}_2)$ 能够构成单纯形迭代的初始基, 因为它产生

基本可行解 $\mathbf{X}_B=(x_1,x_2)^{\mathrm{T}}$. 然而, 如果不用向量表示, 只能利用求解过程来试图找到所有可能的基 (正如上面所演示的那样, 这个例子所有可能基的个数是 3). 使用试错法的困难在于它不适于自动计算. 在一个具有几千个变量和约束的典型的线性规划问题中, 计算机的使用是必需的, 而试错法并不实用, 因为它的计算量非常巨大. 为了解决这个问题, 单纯形法总是以一个单位矩阵 $\mathbf{B}=\mathbf{I}$ 开始其迭代. 为什么要从 $\mathbf{B}=\mathbf{I}$ 开始呢? 因为它总能给出一个可行的初始基本解 (只要方程的右端向量是非负的). 可以从图 13.3 中看到这一结果 (通过画出 $\mathbf{B}=\mathbf{I}$ 的向量, 并注意到它们与横轴和纵轴相重合), 因此, 总能保证初始基本解的可行性.

如果原问题所有的约束均是 $\leqslant$ 类型, 则基 $\mathbf{B}=\mathbf{I}$ 自动构成线性规划等式约束的一部分. 在其他的情况下, 可以在所需要的地方简单地增加单位向量. 这正是人工变量要完成的工作 (见 3.4 节). 然后在目标函数中惩罚这些变量以使得它们在最终的结果中取零值.

习题 13.1B

1. 在下列方程组中, (a) 和 (b) 有唯一的 (基本) 解, (c) 有无穷解, (d) 没有解. 说明如何用图形的向量表示来验证这些结果. 通过这个练习, 陈述导致线性方程组有唯一解、无穷解和无解情况下向量线性相关或线性无关的一般条件.

   (a) $\begin{aligned} x_1+3x_2&=2\\ 3x_1+x_2&=3\end{aligned}$　　(b) $\begin{aligned} 2x_1+3x_2&=1\\ 2x_1-x_2&=2\end{aligned}$

   (c) $\begin{aligned} 2x_1+6x_2&=4\\ x_1+3x_2&=2\end{aligned}$　　(d) $\begin{aligned} 2x_1-4x_2&=2\\ -x_1+2x_2&=1\end{aligned}$

2. 用向量从图形上确定下列方程组解的类型: 唯一解、无穷解或无解. 对于唯一解的情况, 用向量表示的方式 (不用解代数方程的方式) 指出 x_1 和 x_2 的取值为正、为零还是为负.

   (a) $\begin{pmatrix}5&4\\1&-3\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix}=\begin{pmatrix}1\\1\end{pmatrix}$　　*(b) $\begin{pmatrix}2&-2\\1&3\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix}=\begin{pmatrix}1\\3\end{pmatrix}$

   (c) $\begin{pmatrix}2&4\\1&3\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix}=\begin{pmatrix}-2\\-1\end{pmatrix}$　　*(d) $\begin{pmatrix}2&4\\1&2\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix}=\begin{pmatrix}6\\3\end{pmatrix}$

   (e) $\begin{pmatrix}-2&4\\1&-2\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix}=\begin{pmatrix}2\\1\end{pmatrix}$　　*(f) $\begin{pmatrix}1&-2\\0&0\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix}=\begin{pmatrix}1\\1\end{pmatrix}$

3. 考虑下面的方程组:

$$\begin{pmatrix}1\\2\\3\end{pmatrix}x_1+\begin{pmatrix}0\\2\\1\end{pmatrix}x_2+\begin{pmatrix}1\\4\\2\end{pmatrix}x_3+\begin{pmatrix}2\\0\\0\end{pmatrix}x_4=\begin{pmatrix}3\\4\\2\end{pmatrix}$$

   确定下列的任意一种组合是否构成基.

   *(a) $(\mathbf{P}_1,\mathbf{P}_2,\mathbf{P}_3)$　　(b) $(\mathbf{P}_1,\mathbf{P}_2,\mathbf{P}_4)$

   (c) $(\mathbf{P}_2,\mathbf{P}_3,\mathbf{P}_4)$　　*(d) $(\mathbf{P}_1,\mathbf{P}_2,\mathbf{P}_3,\mathbf{P}_4)$

4. 判断正误.

(a) 如果 $\boldsymbol{B}$ 非奇异, 则方程组 $\boldsymbol{BX}=\boldsymbol{b}$ 有唯一解.

(b) 如果 $\boldsymbol{B}$ 奇异且 $\boldsymbol{b}$ 与 $\boldsymbol{B}$ 线性无关, 则方程组 $\boldsymbol{BX}=\boldsymbol{b}$ 无解.

(c) 如果 $\boldsymbol{B}$ 奇异且与 $\boldsymbol{b}$ 线性相关, 则方程组 $\boldsymbol{BX}=\boldsymbol{b}$ 有无穷解.

13.1.2 广义单纯形表的矩阵表示形式

本节用矩阵方法研究一般单纯形表. 这种表示方法将是本章内容的基础.

考虑等式形式的线性规划:

$$\max\ z=\boldsymbol{CX},\quad \text{s.t. } \boldsymbol{AX}=\boldsymbol{b},\ \boldsymbol{X}\geqslant \boldsymbol{0}$$

问题可以等价地表示成

$$\begin{pmatrix}1 & -\boldsymbol{C}\\ \boldsymbol{0} & \boldsymbol{A}\end{pmatrix}\begin{pmatrix}z\\ \boldsymbol{X}\end{pmatrix}=\begin{pmatrix}0\\ \boldsymbol{b}\end{pmatrix}$$

假定 $\boldsymbol{B}$ 是方程组 $\boldsymbol{AX}=\boldsymbol{b}$, $\boldsymbol{X}\geqslant\boldsymbol{0}$ 的可行基, 并且令 $\boldsymbol{X}_B$ 是相应的基变量, $\boldsymbol{C}_B$ 是对应的目标系数向量, 则相应的解可以由下列方式计算 (分块矩阵求逆的方法由附录 D.2.7 给出):

$$\begin{pmatrix}z\\ \boldsymbol{X}_B\end{pmatrix}=\begin{pmatrix}1 & -\boldsymbol{C}_B\\ \boldsymbol{0} & \boldsymbol{B}\end{pmatrix}^{-1}\begin{pmatrix}0\\ \boldsymbol{b}\end{pmatrix}=\begin{pmatrix}1 & \boldsymbol{C}_B\boldsymbol{B}^{-1}\\ \boldsymbol{0} & \boldsymbol{B}^{-1}\end{pmatrix}\begin{pmatrix}0\\ \boldsymbol{b}\end{pmatrix}=\begin{pmatrix}\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{b}\\ \boldsymbol{B}^{-1}\boldsymbol{b}\end{pmatrix}$$

矩阵形式的一般单纯形表可由原标准方程得到, 其格式如下:

$$\begin{pmatrix}1 & \boldsymbol{C}_B\boldsymbol{B}^{-1}\\ \boldsymbol{0} & \boldsymbol{B}^{-1}\end{pmatrix}\begin{pmatrix}1 & -\boldsymbol{C}\\ \boldsymbol{0} & \boldsymbol{A}\end{pmatrix}\begin{pmatrix}z\\ \boldsymbol{X}\end{pmatrix}=\begin{pmatrix}1 & \boldsymbol{C}_B\boldsymbol{B}^{-1}\\ \boldsymbol{0} & \boldsymbol{B}^{-1}\end{pmatrix}\begin{pmatrix}0\\ \boldsymbol{b}\end{pmatrix}$$

矩阵相乘, 产生如下方程:

$$\begin{pmatrix}1 & \boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{A}-\boldsymbol{C}\\ \boldsymbol{0} & \boldsymbol{B}^{-1}\boldsymbol{A}\end{pmatrix}\begin{pmatrix}z\\ \boldsymbol{X}\end{pmatrix}=\begin{pmatrix}\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{b}\\ \boldsymbol{B}^{-1}\boldsymbol{b}\end{pmatrix}$$

已知向量 $\boldsymbol{P}_j$ 是矩阵 $\boldsymbol{A}$ 的第 j 列, 对应的变量是 x_j, 相应的单纯形表具有如下形式:

基	x_j	解
z	$\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{P}_j-c_j$	$\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{b}$
$\boldsymbol{X}_B$	$\boldsymbol{B}^{-1}\boldsymbol{P}_j$	$\boldsymbol{B}^{-1}\boldsymbol{b}$

事实上, 上表与第 3 章所述的单纯形表相同 (见习题 13.1C 的第 5 题), 其计算与 4.2.4 节中原始-对偶方法相同. 这个表的重要性质是, 从一张表变化到下一张表时只是逆矩阵 $\boldsymbol{B}^{-1}$ 发生了变化, 一旦 $\boldsymbol{B}^{-1}$ 已知, 便能够构造出整个单纯形表. 这一点是重要的, 因为任意一张单纯形表的计算舍入误差可以通过控制 $\boldsymbol{B}^{-1}$ 的精确度来控制. 这个结果是 7.2 节介绍的修正单纯形法的基础.

例 13.1-3

考虑如下线性规划：

$$\begin{aligned} \max\ & z = x_1 + 4x_2 + 7x_3 + 5x_4 \\ \text{s.t.}\ & 2x_1 + x_2 + 2x_3 + 4x_4 = 10 \\ & 3x_1 - x_2 - 2x_3 + 6x_4 = 5 \\ & x_1,\ x_2,\ x_3,\ x_4 \geqslant 0 \end{aligned}$$

用基 $\boldsymbol{B} = (\boldsymbol{P}_1, \boldsymbol{P}_2)$ 构造完整的单纯形表.

已知 $\boldsymbol{B} = (\boldsymbol{P}_1, \boldsymbol{P}_2)$, 则 $\boldsymbol{X}_B = (x_1, x_2)^{\mathrm{T}}$ 和 $\boldsymbol{C}_B = (1, 4)$. 因此,

$$\boldsymbol{B}^{-1} = \begin{pmatrix} 2 & 1 \\ 3 & -1 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{5} & \frac{1}{5} \\ \frac{3}{5} & -\frac{2}{5} \end{pmatrix}$$

于是可得

$$\boldsymbol{X}_B = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \boldsymbol{B}^{-1}\boldsymbol{b} = \begin{pmatrix} \frac{1}{5} & \frac{1}{5} \\ \frac{3}{5} & -\frac{2}{5} \end{pmatrix} \begin{pmatrix} 10 \\ 5 \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$$

为计算单纯形表的约束列, 我们有

$$\boldsymbol{B}^{-1}(\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3, \boldsymbol{P}_4) = \begin{pmatrix} \frac{1}{5} & \frac{1}{5} \\ \frac{3}{5} & -\frac{2}{5} \end{pmatrix} \begin{pmatrix} 2 & 1 & 2 & 4 \\ 3 & -1 & -2 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 2 \\ 0 & 1 & 2 & 0 \end{pmatrix}$$

接下来, 计算目标行：

$$\boldsymbol{C}_B(\boldsymbol{B}^{-1}(\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3, \boldsymbol{P}_4)) - \boldsymbol{C} = (1, 4) \begin{pmatrix} 1 & 0 & 0 & 2 \\ 0 & 1 & 2 & 0 \end{pmatrix} - (1, 4, 7, 5) = (0, 0, 1, -3)$$

最后, 计算目标函数值：

$$z = \boldsymbol{C}_B \boldsymbol{B}^{-1} \boldsymbol{b} = \boldsymbol{C}_B \boldsymbol{X}_B = (1, 4) \begin{pmatrix} 3 \\ 4 \end{pmatrix} = 19$$

因此, 完整的单纯形表如下所示.

基	x_1	x_2	x_3	x_4	解
z	0	0	1	−3	19
x_1	1	0	0	2	3
x_2	0	1	2	0	4

从这个例子得到的主要结论是, 一旦逆 $\boldsymbol{B}^{-1}$ 已知, 就能够从 $\boldsymbol{B}^{-1}$ 和问题的原始数据构造出完整的单纯形表.

习题 13.1C

*1. 在例 13.1-3 中, 考虑 $B = (P_3, P_4)$. 证明相应的基本解是可行的, 然后构造出相应的单纯形表.

2. 考虑如下线性规划:

$$\begin{aligned} \max\ & z = 5x_1 + 12x_2 + 4x_3 \\ \text{s.t.}\ & x_1 + 2x_2 + x_3 + x_4 = 10 \\ & 2x_1 - 2x_2 - x_3 = 2 \\ & x_1,\ x_2,\ x_3,\ x_4 \geqslant 0 \end{aligned}$$

核实下列矩阵是否构成 (可行或不可行) 基: (P_1, P_2), (P_2, P_3), (P_3, P_4).

3. 在下面的线性规划中, 对应于 $X_B = (x_1, x_2, x_5)^T$, 计算出完整的单纯形表.

$$\begin{aligned} \min\ & z = 2x_1 + x_2 \\ \text{s.t.}\ & 3x_1 + x_2 - x_3 = 3 \\ & 4x_1 + 3x_2 - x_4 = 6 \\ & x_1 + 2x_2 + x_5 = 3 \\ & x_1,\ x_2,\ x_3,\ x_4,\ x_5 \geqslant 0 \end{aligned}$$

*4. 下面的单纯形表是某个线性规划的最优单纯形表.

基	x_1	x_2	x_3	x_4	x_5	解
z	0	0	0	3	2	?
x_3	0	0	1	1	−1	2
x_2	0	1	0	1	0	6
x_1	1	0	0	−1	1	2

变量 x_3, x_4, x_5 是原问题的松弛变量. 用矩阵运算的方法重构原线性规划, 然后计算其最优值.

5. 在广义单纯形表中, 假定 $X = (X_{\text{I}}, X_{\text{II}})^T$, 其中 X_{II} 对应于一个典型的 $B = I$ 时的初始基本解 (由松弛变量和/或人工变量组成), 并令 $C = (C_{\text{I}}, C_{\text{II}})$ 和 $A = (D, I)$ 分别是 C 和 A 的划分. 证明: 单纯形表的矩阵形式可简化为如下形式, 而这正是第 3 章所采用的单纯形表形式.

基	X_{I}	X_{II}	解
z	$C_B B^{-1} D - C_{\text{I}}$	$C_B B^{-1} - C_{\text{II}}$	$C_B B^{-1} b$
X_B	$B^{-1} D$	B^{-1}	$B^{-1} b$

13.2 修正单纯形法

13.1.1 节证明了线性规划的最优解总是与一个基本 (可行) 解相关. 单纯形法通过选择一个可行基 B 寻找一个最合适的初始点, 然后移动到下一个基 B_{next}, 产生一个更好的 (至少是相等的) 目标函数值. 持续这一过程, 最终达到最优基.

修正单纯形法的迭代步骤与第 3 章提出的*表格形式*的单纯形法完全相同. 主要差别是, 修正单纯形法的计算是基于矩阵运算而不再是行运算. 使用矩阵代数可以通过控制计算 $\boldsymbol{B}^{-1}$ 的精度来减轻机器舍入误差的不利影响. 得出这个结论是因为, 正像 13.1.2 节所说明的那样, 整个单纯形表的计算可由*原始*数据和当前的 $\boldsymbol{B}^{-1}$ 得到. 而在第 3 章的表格单纯形法中, 每一张表都是由前一张表直接生成的, 这导致了更大的舍入误差问题.

13.2.1 最优性条件与可行性条件的建立

一般线性规划问题能够写成如下的表达形式:

$$\max \text{或} \min\ z=\sum_{j=1}^{n} c_j x_j,\quad \text{s.t.}\ \sum_{j=1}^{n} \boldsymbol{P}_j x_j=\boldsymbol{b},\ x_j \geqslant 0,\ j=1,2,\cdots,n$$

对于已知的基向量 $\boldsymbol{X}_B$, 以及它所对应的基 $\boldsymbol{B}$ 和目标向量 $\boldsymbol{C}_B$, 13.1.2 节给出的广义单纯形表表明, 任意的单纯形迭代都可以表示成下列方程:

$$z+\sum_{j=1}^{n}(z_j-c_j)\,x_j=\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{b}$$

$$(\boldsymbol{X}_B)_i+\sum_{j=1}^{n}\left(\boldsymbol{B}^{-1}\boldsymbol{P}_j\right)_i x_j=\left(\boldsymbol{B}^{-1}\boldsymbol{b}\right)_i$$

z_j-c_j, 即 x_j 的简约费用 (见 4.3.2 节), 定义为

$$z_j-c_j=\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{P}_j-c_j$$

用记号 $(\boldsymbol{V})_i$ 表示向量 $\boldsymbol{V}$ 的第 i 个元素.

最优性条件 (optimality condition) 从上面给出的 z 行方程可以看出, 非基变量 x_j 由当前的零值增加时, 将改善 z 的当前值 $(\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{b})$: 对于求极大值, 只要 z_j-c_j 严格负; 对于求极小值, 只要 z_j-c_j 严格正. 否则, x_j 不能改进解, 并且必须保持非基变量为零. 尽管选择任何满足上述给定条件的非基变量均能改进当前解, 但单纯形法采用的是单凭经验的方法, 就是在求极大 (求极小) 情况下选择最负的 (最正的) z_j-c_j 所对应的变量, 称为**进基变量** (entering variable).

可行性条件 (feasibility condition) 可以通过检查与第 i 个基变量所对应的约束方程来确定**离基向量** (leaving vector). 特别地, 我们有

$$(\boldsymbol{X}_B)_i+\sum_{j=1}^{n}\left(\boldsymbol{B}^{-1}\boldsymbol{P}_j\right)_i x_j=\left(\boldsymbol{B}^{-1}\boldsymbol{b}\right)_i$$

当最优性条件选择 $\boldsymbol{P}_j$ 作为进基向量, 它相应的变量 x_j 将从零值开始增加. 同时, 其余所有的非基变量仍保持为零值. 因此, 第 i 个约束方程简化为

$$(\boldsymbol{X}_B)_i=\left(\boldsymbol{B}^{-1}\boldsymbol{b}\right)_i-\left(\boldsymbol{B}^{-1}\boldsymbol{P}_j\right)_i x_j$$

此方程表明, 如果 $\left(B^{-1}P_j\right)_i > 0$, 则 x_j 的增加能够使 $(X_B)_i$ 变为负值, 这样会破坏非负性条件: 对所有的 i, $(X_B)_i \geqslant 0$. 因此, 我们有

$$\left(B^{-1}b\right)_i - \left(B^{-1}P_j\right)_i x_j \geqslant 0, \ \forall\, i$$

这个条件产生进基变量 x_j 的最大值, 即

$$x_j = \min_i \left\{ \frac{\left(B^{-1}b\right)_i}{\left(B^{-1}P_j\right)_i} \;\middle|\; \left(B^{-1}P_j\right)_i > 0 \right\}$$

基本解中与最小比值相对应的离基变量成为非基变量, 其取值为零.

习题 13.2A

*1. 考虑下面的线性规划:

$$\max \ z = c_1x_1 + c_2x_2 + c_3x_3 + c_4x_4$$

$$\text{s.t.} \ \ P_1x_1 + P_2x_2 + P_3x_3 + P_4x_4 = b$$

$$x_1, x_2, x_3, x_4 \geqslant 0$$

向量 P_1, P_2, P_3, P_4 如图 13.4 所示. 假定当前用于迭代的基 B 由 P_1 和 P_2 组成.

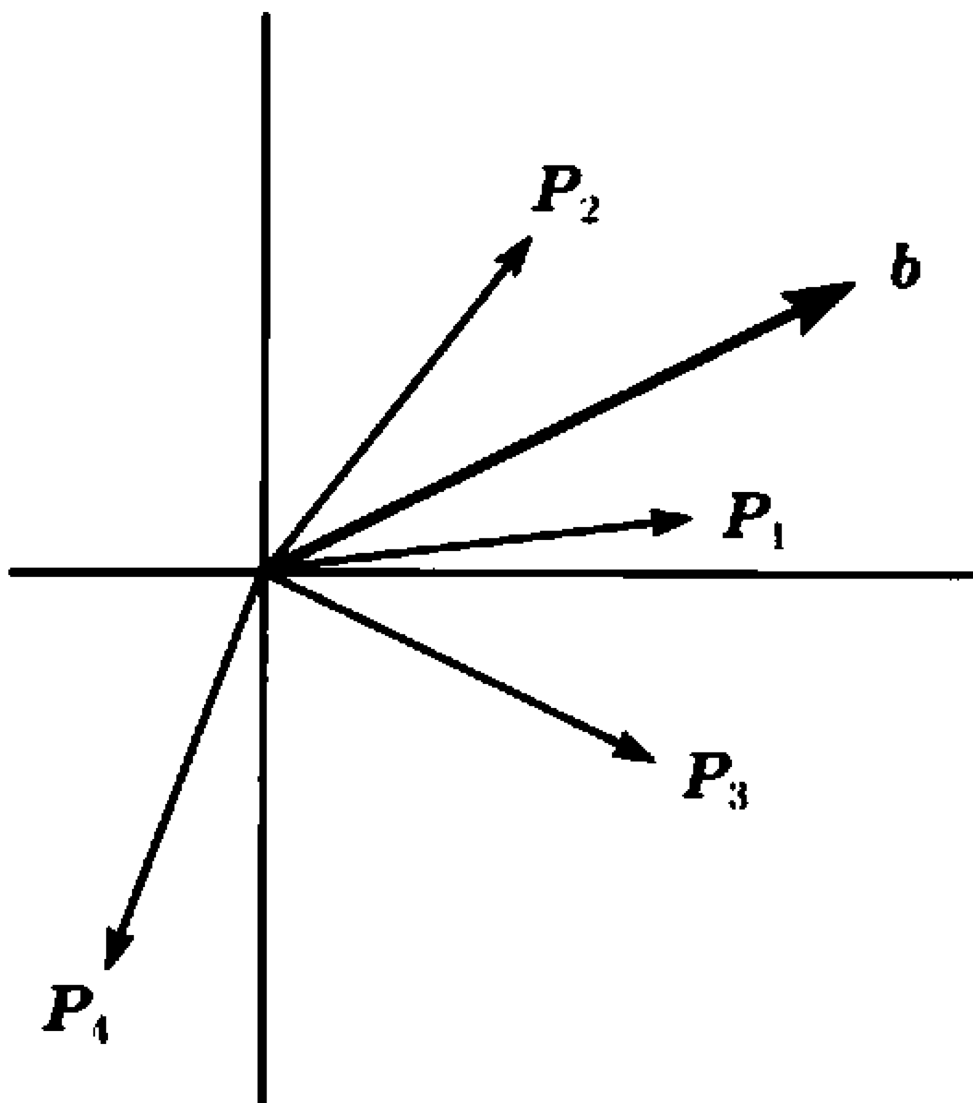


图 13.4 习题 13.2A 第 1 题的向量表示

(a) 如果向量 P_3 进基, 当前两个基向量中的哪一个必须离开, 使得到的基本解是可行的?

(b) 向量 P_4 能成为可行基的一部分吗?

*2. 证明: 在任何单纯形迭代中, $z_j - c_j = 0$ 对于所有的基变量均成立.

3. 证明: 在极大化 (极小化) 线性规划问题中, 若对所有非基变量 x_j 均有 $z_j - c_j > 0(< 0)$, 则最优解是唯一的. 否则, 如果存在一个非基变量 x_j, 使得 $z_j - c_j = 0$, 则问题有另一个最优解.

4. 在一个全部松弛变量为初始基本解的问题中, 证明: 若用矩阵形式表示单纯形表, 其求解过程如 3.3 节所示, 对应的目标方程为

$$z-\sum_{j=1}^{n}c_jx_j=0$$

则在初始单纯形表中将自动地计算出所有变量相应的 z_j-c_j 的值.

5. 使用矩阵形式的单纯形表, 证明: 在一个全部人工变量为初始基本解的问题中, 采用 3.4.1 节的计算过程, (用约束方程) 替换出目标函数中全部的人工变量, 则得到的初始单纯形表将自动地计算出所有变量相应的 z_j-c_j 的值.

6. 考虑一个线性规划, 在此线性规划中, 变量 x_k 无符号限制. 证明: 通过替换 $x_k=x_k^- - x_k^+$, 其中 x_k^- 和 x_k^+ 非负, 则两个变量在不同的最优解中彼此相互替换是不可能的.

***7.** 给出一个一般形式的线性规划, 其方程由 m 个方程和 n 个未知量构成, 确定可由解空间中一个非退化极点 (所有基变量 >0) 到达的邻近极点的最大数目.

8. 在应用单纯形法的可行性条件时, 假定 $x_r=0$ 是一个基变量, x_j 是进基变量, 并且 $(\boldsymbol{B}^{-1}\boldsymbol{P}_j)_r\neq 0$. 证明: 即使 $(\boldsymbol{B}^{-1}\boldsymbol{P}_j)_r$ 为负值, 导出的基本解仍保持可行.

9. 在单纯形法可行性条件的实施过程中, 第一次遇到退化解 (至少一个基变量 $=0$) 的条件是什么? 在下一个迭代中继续得到退化解的条件又是什么? 从数学角度解释你的回答.

***10.** 在退化和非退化的情况下, 极点与基本解的关系是什么? 在给定一个极点且假定非循环的情况下, 最大迭代的数目是多少?

***11.** 考虑线性规划: max $z=\boldsymbol{CX}$, s.t. $\boldsymbol{AX}\leqslant \boldsymbol{b}$, $\boldsymbol{X}\geqslant \boldsymbol{0}$, 其中 $\boldsymbol{b}\geqslant\boldsymbol{0}$. 假定 $\boldsymbol{P}_j$ 是进基向量, 且使 $\boldsymbol{B}^{-1}\boldsymbol{P}_j$ 中至少有一个元素为正.

(a) 如果 $\boldsymbol{P}_j$ 用 $\alpha\boldsymbol{P}_j$ 替换, 其中 α 是一个正的标量, 而且倘若 x_j 仍保持为进基变量, 找出 x_j 的值与 $\boldsymbol{P}_j$ 和 $\alpha\boldsymbol{P}_j$ 之间的关系.

(b) 如果还加上 $\boldsymbol{b}$ 用 $\beta\boldsymbol{b}$ 替换, 其中 β 是一个正的标量, 重解问题 (a).

12. 考虑线性规划

$$\max\ z=\boldsymbol{CX},\quad \text{s.t. } \boldsymbol{AX}\leqslant \boldsymbol{b},\ \boldsymbol{X}\geqslant \boldsymbol{0},\quad \text{其中 } \boldsymbol{b}\geqslant\boldsymbol{0}$$

在得到最优解后, 建议通过减少 x_j 的单位 (资源) 需求, 使非基变量 x_j 成为基变量 (有利的), 其方法是, 对于不同资源, 在原值上乘上 $\frac{1}{\alpha}$, 其中 $\alpha>1$. 因为单位需求减少了, 所以预计 x_j 的单位利润也将减少到它原值的 $\frac{1}{\alpha}$. 这种变化能使 x_j 成为有利可图的变量吗? 从数学上予以解释.

13. 考虑线性规划

$$\max\ z=\boldsymbol{CX},\quad \text{s.t. } (\boldsymbol{A},\boldsymbol{I})\boldsymbol{X}=\boldsymbol{b},\ \boldsymbol{X}\geqslant\boldsymbol{0}$$

定义 $\boldsymbol{X}_B$ 是以 $\boldsymbol{B}$ 为基的基向量, $\boldsymbol{C}_B$ 是相应的目标系数向量. 证明: 如果 $\boldsymbol{C}_B$ 由新的系数 $\boldsymbol{D}_B$ 替换, 对于基向量 $\boldsymbol{X}_B$, 其 z_j-c_j 的值还将保持为零. 这个结果的意义是什么?

13.2.2 修正单纯形算法

13.2.1 节得到了最优性条件和可行性条件, 现在介绍修正单纯形法的计算步骤.

第 0 步 构造一个初始基本可行解, 并令 $\boldsymbol{B}$ 和 $\boldsymbol{C}_B$ 分别是相应的基向量和目标系数向量.

第 1 步 选择一种适当的求逆方法[①] 计算逆 $\boldsymbol{B}^{-1}$.

第 2 步 对于每个非基变量 x_j, 计算

$$z_j - c_j = \boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{P}_j - c_j$$

如果对于所有的非基变量 x_j, 在极大化问题中 $z_j - c_j \geqslant 0$(在极小化问题中 $\leqslant 0$), 则停止计算, 得到最优解

$$\boldsymbol{X}_B = \boldsymbol{B}^{-1}\boldsymbol{b}, \quad z = \boldsymbol{C}_B\boldsymbol{X}_B$$

否则, 应用最优性条件, 对应于求极大 (极小) 值, 把最负 (正) 的 $z_j - c_j$ 所对应的非基变量确定为进基变量.

第 3 步 计算 $\boldsymbol{B}^{-1}\boldsymbol{P}_j$. 如果 $\boldsymbol{B}^{-1}\boldsymbol{P}_j$ 中所有的元素为负值或零, 则停止计算, 问题无有界解. 否则, 计算 $\boldsymbol{B}^{-1}\boldsymbol{b}$. 然后, 对 $\boldsymbol{B}^{-1}\boldsymbol{P}_j$ 中所有的严格正元素, 由可行性条件的定义确定其比值. 相应于最小比值的基变量 x_i 为离基变量.

第 4 步 在当前基 $\boldsymbol{B}$ 中, 通过用进基向量 $\boldsymbol{P}_j$ 替代离基向量 $\boldsymbol{P}_i$ 来构造新的基. 转入第 1 步, 开始新的迭代.

例 13.2-1

用修正单纯形算法求解 Reddy Mikks 模型 (2.1 节). 3.3.2 节中已经使用单纯形表求解过相同的模型. 比较这两种方法, 将看出它们完全一样.

Reddy Mikks 模型的矩阵表示如下:

$$\max \ z = (5,4,0,0,0,0)(x_1,x_2,x_3,x_4,x_5,x_6)^{\mathrm{T}}$$

$$\text{s.t.} \quad \begin{pmatrix} 6 & 4 & 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} = \begin{pmatrix} 24 \\ 6 \\ 1 \\ 2 \end{pmatrix}$$

$$x_1, x_2, x_3, x_4, x_5, x_6 \geqslant 0$$

我们用符号 $\boldsymbol{C} = (c_1, c_2, \cdots, c_6)$ 表示目标函数的系数, 并用 $(\boldsymbol{P}_1, \boldsymbol{P}_2, \cdots, \boldsymbol{P}_6)$ 表示约束方程的列向量. 约束的右端项由向量 $\boldsymbol{b}$ 给出.

---

① 在大多数线性规划的介绍中, 包括本书的前 6 版, 都将求基逆 (见附录 D.2.7) 的乘积结构方法统一到了修正单纯形算法中, 因为乘积结构非常有助于修正单纯算法的计算, 其中在相继的迭代中, 其基恰好有一列不同. 这个细节已在这次的介绍中去掉, 因为它使得算法看起来比真实的算法更复杂. 此外, 乘积结构已经很少用于线性规划代码的开发, 因为它不是为自动计算而设计的, 在自动计算中, 机械舍入误差可能是一个严重的问题. 通常, 一些高级的数值分析方法, 像 LU分解方法, 都可用于矩阵的求逆. (顺便提一句, TORA 矩阵求逆就是基于 LU 分解的.)

在下面的计算中，我们将给出每一步计算的代数式，并直接给出它最终的数值结果而不附加详细的算术运算. 你将发现这样处理有助于说明各步骤的关系.

迭代 0

$$X_{B_0}=(x_3,x_4,x_5,x_6)^{\mathrm{T}},\quad C_{B_0}=(0,0,0,0)$$
$$B_0=(P_3,P_4,P_5,P_6)=I,\quad B_0^{-1}=I$$

因此,

$$X_{B_0}=B_0^{-1}b=(24,6,1,2)^{\mathrm{T}},\quad z=C_{B_0}X_{B_0}=0$$

最优性计算:

$$C_{B_0}B_0^{-1}=(0,0,0,0)$$
$$\{z_j-c_j\}_{j=1,2}=C_{B_0}B_0^{-1}(P_1,P_2)-(c_1,c_2)=(-5,-4)$$

因此, P_1 是进基向量.

可行性计算:

$$X_{B_0}=(x_3,x_4,x_5,x_6)^{\mathrm{T}}=(24,6,1,2)^{\mathrm{T}}$$
$$B_0^{-1}P_1=(6,1,-1,0)^{\mathrm{T}}$$

因此,

$$x_1=\min\left\{\frac{24}{6},\frac{6}{1},-,-\right\}=\min\{4,6,-,-\}=4$$

因而, P_3 成为离基向量.

上述结果可以总结为熟悉的单纯形表格形式. 这种表示将帮助你确信两种方法本质上是相同的. 你将发现, 在随后的迭代中, 它有益于开发类似的表格.

基	x_1	x_2	x_3	x_4	x_5	x_6	解
z	−5	−4	0	0	0	0	0
x_3	6						24
x_4	1						6
x_5	−1						1
x_6	0						2

迭代 1

$$X_{B_1}=(x_1,x_4,x_5,x_6)^{\mathrm{T}},\quad C_{B_1}=(5,0,0,0)$$
$$B_1=(P_1,P_4,P_5,P_6)=\begin{pmatrix}6&0&0&0\\1&1&0&0\\-1&0&1&0\\0&0&0&1\end{pmatrix}$$

采用适当的求逆的方法 (见附录 D.2.7, 特别是乘积结构方法), 得到的逆矩阵为

$$\boldsymbol{B}_1^{-1}=\begin{pmatrix} \frac{1}{6} & 0 & 0 & 0 \\ -\frac{1}{6} & 1 & 0 & 0 \\ \frac{1}{6} & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

因此,

$$\boldsymbol{X}_{B_1}=\boldsymbol{B}_1^{-1}\boldsymbol{b}=(4,2,5,2)^{\mathrm{T}}, \quad z=\boldsymbol{C}_{B_1}\boldsymbol{X}_{B_1}=20$$

最优性计算:

$$\boldsymbol{C}_{B_1}\boldsymbol{B}_1^{-1}=\left(\frac{5}{6},0,0,0\right)$$

$$\{z_j-c_j\}_{j=2,3}=\boldsymbol{C}_{B_1}\boldsymbol{B}_1^{-1}(\boldsymbol{P}_2,\boldsymbol{P}_3)-(c_2,c_3)=\left(-\frac{2}{3},\frac{5}{6}\right)$$

因此, $\boldsymbol{P}_2$ 是进基向量.

可行性计算:

$$\boldsymbol{X}_{B_1}=(x_1,x_4,x_5,x_6)^{\mathrm{T}}=(4,2,5,2)^{\mathrm{T}}$$

$$\boldsymbol{B}_1^{-1}\boldsymbol{P}_2=\left(\frac{2}{3},\frac{4}{3},\frac{5}{3},1\right)^{\mathrm{T}}$$

因此,

$$x_2=\min\left\{\frac{4}{\frac{2}{3}},\frac{2}{\frac{4}{3}},\frac{5}{\frac{5}{3}},\frac{2}{1}\right\}=\min\left\{6,\frac{\mathbf{3}}{\mathbf{2}},3,2\right\}=\frac{3}{2}$$

因而, P_4 成为离基向量. (你将发现前面在迭代 0 中用单纯形表来概括结果是很有帮助的.)

迭代 2

$$\boldsymbol{X}_{B_2}=(x_1,x_2,x_5,x_6)^{\mathrm{T}}, \quad \boldsymbol{C}_{B_2}=(5,4,0,0)$$

$$\boldsymbol{B}_2=(\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{P}_5,\boldsymbol{P}_6)=\begin{pmatrix} 6 & 4 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ -1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

因此,

$$\boldsymbol{B}_2^{-1}=\begin{pmatrix} \frac{1}{4} & -\frac{1}{2} & 0 & 0 \\ -\frac{1}{8} & \frac{3}{4} & 0 & 0 \\ \frac{3}{8} & -\frac{5}{4} & 1 & 0 \\ \frac{1}{8} & -\frac{3}{4} & 0 & 1 \end{pmatrix}$$

从而,

$$X_{B_2}=B_2^{-1}b=\left(3,\frac{3}{2},\frac{5}{2},\frac{1}{2}\right)^{\mathrm{T}},\quad z=C_{B_2}X_{B_2}=21$$

最优性计算:

$$C_{B_2}B_2^{-1}=\left(\frac{3}{4},\frac{1}{2},0,0\right)$$

$$\{z_j-c_j\}_{j=3,4}=C_{B_2}B_2^{-1}(P_3,P_4)-(c_3,c_4)=\left(\frac{3}{4},\frac{1}{2}\right)$$

因此, X_{B_2} 是最优解, 计算结束.

最优解的概括:

$$x_1=3,\quad x_2=1.5,\quad z=21$$

习题 13.2B

1. 在例 13.2-1 中, 概述 3.3 节表格形式中迭代 1 的数据.
2. 用修正单纯形方法求解下列线性规划:

(a)
$$\begin{aligned}\max\ & z=6x_1-2x_2+3x_3\\ \text{s.t.}\ & 2x_1-x_2+2x_3\leqslant 2\\ & x_1\qquad +4x_3\leqslant 4\\ & x_1,x_2,x_3\geqslant 0\end{aligned}$$

*(b)
$$\begin{aligned}\max\ & z=2x_1+x_2+2x_3\\ \text{s.t.}\ & 4x_1+3x_2+8x_3\leqslant 12\\ & 4x_1+x_2+12x_3\leqslant 8\\ & 4x_1-x_2+3x_3\leqslant 8\\ & x_1,x_2,x_3\geqslant 0\end{aligned}$$

(c)
$$\begin{aligned}\min\ & z=2x_1+x_2\\ \text{s.t.}\ & 3x_1+x_2=3\\ & 4x_1+3x_2\geqslant 6\\ & x_1+2x_2\leqslant 3\\ & x_1,x_2\geqslant 0\end{aligned}$$

(d)
$$\begin{aligned}\min\ & z=5x_1-4x_2+6x_3+8x_4\\ \text{s.t.}\ & x_1+7x_2+3x_3+7x_4\leqslant 46\\ & 3x_1-x_2+x_3+2x_4\leqslant 20\\ & 2x_1+3x_2-x_3+x_4\geqslant 18\\ & x_1,x_2,x_3,x_4\geqslant 0\end{aligned}$$

3. 用修正单纯形法求解下面的线性规划, 已知可行的初始基向量 $X_{B_0}=(x_2,x_4,x_5)^{\mathrm{T}}$.

$$\begin{aligned}\min\ & z=7x_2+11x_3-10x_4+26x_6\\ \text{s.t.}\ & x_2-x_3+x_5+x_6=6\\ & x_2-x_3+x_4+3x_6=8\\ & x_1+x_2-3x_3+x_4+x_5=12\\ & x_1,x_2,x_3,x_4,x_5,x_6\geqslant 0\end{aligned}$$

4. 用两阶段修正单纯形方法求解下面的问题:
   (a) 第 2-c 题.　　(b) 第 2-d 题.
   (c) 第 3 题 (不考虑给出的初始 X_{B_0}).
5. *修正对偶单纯形法*. 修正对偶单纯形法 (使用矩阵运算) 的计算步骤总结如下:

第 0 步　令 $B_0=I$ 是初始基, 而且 X_{B_0} 的元素至少有一个取负值 (不可行).

第 1 步 计算 $X_B = B^{-1}b$ 为基变量的当前值. 选择最负的 x_r 作为离基变量. 如果 X_B 的所有元素均非负, 停止计算, 当前解是可行的 (并且是最优的).

第 2 步 (a) 对于所有的非基变量 x_j, 计算 $z_j - c_j = C_B B^{-1} P_j - c_j$.

(b) 对于所有的非基变量 x_j, 对应于离基变量 x_r 的行, 计算约束系数 $\left(B^{-1}P_j\right)_r$.

(c) 进基变量与下面的比值有关:

$$\theta = \min_j \left\{ \left| \frac{z_j - c_j}{\left(B^{-1}P_j\right)_r} \right|,\ \left(B^{-1}P_j\right)_r < 0 \right\}$$

如果所有的 $\left(B^{-1}P_j\right)_r \geqslant 0$, 则可行解不存在.

第 3 步 交换进基向量和离基向量 (P_j 与 P_r) 得到新的基. 计算新的逆矩阵, 并转到第 1 步.

应用这种方法求解下面的问题:

$$\begin{aligned}
\min\ \ & z = 3x_1 + 2x_2 \\
\text{s.t.}\ \ & 3x_1 + \ \ x_2 \geqslant 3 \\
& 4x_1 + 3x_2 \geqslant 6 \\
& \ \ x_1 + 2x_2 \leqslant 3 \\
& \ \ x_1, x_2 \geqslant 0
\end{aligned}$$

13.3 有界变量算法

在线性规划模型中, 可能会明确地给出变量的上界和下界. 例如, 在生产设备中, 可以用下界和上界表示某些产品的最小和最大需求. 有界变量还经常用在求解整数规划的分支定界算法 (见 8.3.1 节) 中.

有界变量算法的计算效率高, 因为它隐式地考虑到了变量的边界. 先考虑下界情况, 因为它更简单些. 假设 $X \geqslant L$, 我们可以在整个问题中运用代换

$$X = L + X', \quad X' \geqslant 0$$

并求解以 X' 为变量的问题 (X' 的下界现在为 0). 原来的 X 由反代换得到, 这是合理的, 因为对所有的 $X' \geqslant 0$, 它保证 $X = X' + L$ 保持非负.

下面考虑上界约束 $X \leqslant U$. 直接代换的概念 (也就是 $X = U - X''$, $X'' \geqslant 0$) 并不正确, 因为逆代换 $X = U - X''$ 并不能保证 X 保持非负. 因此, 需要运用不同的处理方法.

定义上有界的线性规划模型如下:

$$\max\ z = \{CX \mid (A, I)X = b,\ 0 \leqslant X \leqslant U\}$$

有界算法仅使用了约束 $(A, I)X = b$, $X \geqslant 0$, 而通过修改单纯形法的可行性条件来隐式地说明 $X \leqslant U$.

令 $\boldsymbol{X}_B = \boldsymbol{B}^{-1}\boldsymbol{b}$ 是 $(\boldsymbol{A}, \boldsymbol{I})\boldsymbol{X} = \boldsymbol{b}$, $\boldsymbol{X} \geqslant \boldsymbol{0}$ 当前的基本可行解, 并按照 (常规的) 最优性条件, $\boldsymbol{P}_j$ 是进基向量. 若所有其他的非基变量为零, 则第 i 个基变量的约束方程可以写成

$$(\boldsymbol{X}_B)_i = \left(\boldsymbol{B}^{-1}\boldsymbol{b}\right)_i - \left(\boldsymbol{B}^{-1}\boldsymbol{P}_j\right)_i x_j$$

当进基变量 x_j 由零开始增加时, $(\boldsymbol{X}_B)_i$ 将根据 $\left(\boldsymbol{B}^{-1}\boldsymbol{P}_j\right)_i$ 取负值或取正值而增大或减小. 因此, 在确定进基变量 x_j 中, 有 3 个条件必须满足.

(1) 基变量 $(\boldsymbol{X}_B)_i$ 保持非负, 也就是 $(\boldsymbol{X}_B)_i \geqslant 0$.

(2) 基变量 $(\boldsymbol{X}_B)_i$ 不超过它的上界, 也就是 $(\boldsymbol{X}_B)_i \leqslant (\boldsymbol{U}_B)_i$, 其中 $\boldsymbol{U}_B$ 由 $\boldsymbol{U}$ 中与 $\boldsymbol{X}_B$ 对应的元素组成.

(3) 进基变量 x_j 不能取到大于其上界的值, 也就是 $x_j \leqslant u_j$, 其中 u_j 是 $\boldsymbol{U}$ 的第 j 个元素.

第一个条件 $(\boldsymbol{X}_B)_i \geqslant 0$ 需要

$$\left(\boldsymbol{B}^{-1}\boldsymbol{b}\right)_i - \left(\boldsymbol{B}^{-1}\boldsymbol{P}_j\right)_i x_j \geqslant 0$$

这个条件满足仅当

$$x_j \leqslant \theta_1 = \min_i \left\{ \frac{\left(\boldsymbol{B}^{-1}\boldsymbol{b}\right)_i}{\left(\boldsymbol{B}^{-1}\boldsymbol{P}_j\right)_i} \;\middle|\; \left(\boldsymbol{B}^{-1}\boldsymbol{P}_j\right)_i > 0 \right\}$$

这个条件与常规单纯形法的可行性条件完全相同.

接下来, 条件 $(\boldsymbol{X}_B)_i \leqslant (\boldsymbol{U}_B)_i$ 指明

$$\left(\boldsymbol{B}^{-1}\boldsymbol{b}\right)_i - \left(\boldsymbol{B}^{-1}\boldsymbol{P}_j\right)_i x_j \leqslant (\boldsymbol{U}_B)_i$$

它被满足仅当

$$x_j \leqslant \theta_2 = \min_i \left\{ \frac{\left(\boldsymbol{B}^{-1}\boldsymbol{b}\right)_i - (\boldsymbol{U}_B)_i}{\left(\boldsymbol{B}^{-1}\boldsymbol{P}_j\right)_i} \;\middle|\; \left(\boldsymbol{B}^{-1}\boldsymbol{P}_j\right)_i < 0 \right\}$$

结合这 3 个限制, x_j 在满足这 3 个条件的情况下进入解, 也就是,

$$x_j = \min\{\theta_1, \theta_2, \mu_j\}$$

对于下一步迭代, 基的变化依赖于 x_j 在 θ_1, θ_2, μ_j 的水平上是否进入解. 假定 $(\boldsymbol{X}_B)_r$ 是离基变量, 则有下列规则.

(1) $x_j = \theta_1$: $(\boldsymbol{X}_B)_r$ 离开基本解 (变为非基), 其取值为零. 用常规的单纯形法得到新的迭代, 此时 x_j 和 $(\boldsymbol{X}_B)_r$ 分别作为进基和离基变量.

(2) $x_j = \theta_2$: $(\boldsymbol{X}_B)_r$ 变为非基变量, 并取它的上界值. 类似于 $x_j = \theta_1$ 的情况得到新的迭代, 但需要作一项修正, 就是要考虑到这样一个事实: $(\boldsymbol{X}_B)_r$ 变为非基

并取它的上界值. 因为 θ_1 和 θ_2 需要所有的非基变量取零值 (请你自己证明的确如此), 我们必须将达到上界的新非基 $(\boldsymbol{X}_B)_r$ 转换成一个取零值的非基变量. 这一要求可以用代换 $(\boldsymbol{X}_B)_r = (\boldsymbol{U}_B)_r - (\boldsymbol{X}'_B)_r$ 来达到, 其中 $(\boldsymbol{X}'_B)_r \geqslant 0$. 在计算新的基之前还是之后做代换, 这是无关紧要的.

(3) $x_j = u_j$: 基向量 $\boldsymbol{X}_B$ 保持不变, 因为 $x_j = u_j$ 是在迫使当前基变量中任何一个变量达到它的下界 (= 0) 或上界之前停止. 这意味着 x_j 将保持为非基变量但达到其上界. 下面将要介绍由代换 $x_j = u_j - x'_j$ 生成的新迭代.

θ_1、θ_2 和 u_j 间的联系可随意解除, 但最好尽可能执行规则 $x_j = u_j$, 因为它只要较少的计算.

代换 $x_j = u_j - x'_j$ 将原始数据 c_j, $\boldsymbol{P}_j$ 变成 $c'_j = -c_j$, $\boldsymbol{P}'_j = -\boldsymbol{P}_j$, 将 $\boldsymbol{b}$ 变成 $\boldsymbol{b}' = \boldsymbol{b} - u_j\boldsymbol{P}_j$. 这意味着如果使用修正单纯形方法, 所有的计算 (例如, $\boldsymbol{B}^{-1}$, $\boldsymbol{X}_B$, $z_j - c_j$) 将使用每次迭代中 $\boldsymbol{C}$, $\boldsymbol{A}$, $\boldsymbol{b}$ 的更新值 (进一步的细节请见习题 13.3A 的第 5 题).

例 13.3-1

用上有界算法求解以下线性规划模型[①]:

$$
\begin{aligned}
\max\ & z = 3x_1 + 5y + 2x_3 \\
\text{s.t.}\ & x_1 + y + 2x_3 \leqslant 14 \\
& 2x_1 + 4y + 3x_3 \leqslant 43 \\
& 0 \leqslant x_1 \leqslant 4,\ 7 \leqslant y \leqslant 10,\ 0 \leqslant x_3 \leqslant 3
\end{aligned}
$$

对 y 的下界用代换 $y = x_2 + 7$, 其中, $0 \leqslant x_2 \leqslant 10 - 7 = 3$.

为了避免被计算的细节 "转移思路", 我们并不用修正单纯形方法来完成计算. 相反, 我们将用紧凑的表格形式计算. 习题 13.3A 的第 5 题至第 7 题使用算法的修正版.

迭代 0

基	x_1	x_2	x_3	x_4	x_5	解
z	-3	-5	-2	0	0	35
x_4	1	1	2	1	0	7
x_5	2	4	3	0	1	15

我们有 $\boldsymbol{B} = \boldsymbol{B}^{-1} = \boldsymbol{I}$, $\boldsymbol{X}_B = (x_4, x_5)^{\mathrm{T}} = \boldsymbol{B}^{-1}\boldsymbol{b} = (7, 15)^{\mathrm{T}}$. 已知 x_2 是进基变量 $(z_2 - c_2 = -5)$, 我们得到

$$
\boldsymbol{B}^{-1}\boldsymbol{P}_2 = (1, 4)^{\mathrm{T}}
$$

① 可以用 TORA 的 Linear Programming ⇒ [illegible] ⇒ [illegible] ⇒ [illegible] ⇒ Bounded simplex 来生成相应的单纯形迭代 (文件 toraEx13.3-1.txt).

于是得

$$\theta_1 = \min\left\{\frac{7}{1}, \frac{15}{4}\right\} = 3.75, \text{对应于 } x_5$$
$$\theta_2 = \infty \ (\text{因为 } \boldsymbol{B}^{-1}\boldsymbol{P}_2 > \boldsymbol{0})$$

下面给出进基变量的上界, 即 $x_2 \leqslant 3$, 其依据如下:

$$x_2 = \min\{3.75, \infty, 3\} = 3\ (= u_2)$$

因为 x_2 在其上界处 $(= u_2 = 3)$ 进基, $\boldsymbol{X}_B$ 保持不变, 故 x_2 成为取值为其上界的非基变量. 我们用代换 $x_2 = 3 - x_2'$ 来获得如下新的单纯形表:

基	x_1	x_2'	x_3	x_4	x_5	解
z	-3	5	-2	0	0	50
x_4	1	-1	2	1	0	4
x_5	2	-4	3	0	1	3

代换确实改变了原单纯形表的右端项, 由 $\boldsymbol{b} = (7, 15)^{\mathrm{T}}$ 变到了 $\boldsymbol{b}' = (4, 3)^{\mathrm{T}}$. 在进一步的计算中应当考虑这一改变.

迭代 1 进基变量是 x_1. 基向量 $\boldsymbol{X}_B$ 和 $\boldsymbol{B}^{-1}(= \boldsymbol{I})$ 与迭代 0 相同. 于是,

$$\boldsymbol{B}^{-1}\boldsymbol{P}_1 = (1, 2)^{\mathrm{T}}$$

$$\theta_1 = \min\left\{\frac{4}{1}, \frac{3}{2}\right\} = 1.5, \text{对应于基变量 } x_5$$

$$\theta_2 = \infty \ (\text{因为 } \boldsymbol{B}^{-1}\boldsymbol{P}_1 > \boldsymbol{0})$$

因此,

$$x_1 = \min\{1.5, \infty, 4\} = 1.5\ (= \theta_1)$$

所以, 进基变量 x_1 成为基变量, 离基变量 x_5 变成取值为零的非基变量, 由此得到

基	x_1	x_2'	x_3	x_4	x_5	解
z	0	-1	$\frac{5}{2}$	0	$\frac{3}{2}$	$\frac{109}{2}$
x_4	0	1	$\frac{1}{2}$	1	$-\frac{1}{2}$	$\frac{5}{2}$
x_1	1	-2	$\frac{3}{2}$	0	$\frac{1}{2}$	$\frac{3}{2}$

迭代 2 新的逆矩阵是

$$\boldsymbol{B}^{-1} = \begin{pmatrix} 1 & -\frac{1}{2} \\ 0 & \frac{1}{2} \end{pmatrix}$$

因此,

$$X_B = (x_4, x_1)^T = B^{-1}b' = \left(\frac{5}{2}, \frac{3}{2}\right)^T$$

其中 $b' = (4,3)^T$ 是由迭代 0 计算的. 选择 x_2' 作为进基变量, 并注意, $P_2' = -P_2$, 我们得到

$$B^{-1}P_2' = (1,-2)^T$$

因此,

$$\theta_1 = \min\left\{\frac{\frac{5}{2}}{1}, -\right\} = 2.5, \text{对应于基变量 } x_4$$

$$\theta_2 = \min\left\{-, \frac{\frac{3}{2}-4}{-2}\right\} = 1.25, \text{对应于基变量 } x_1$$

则有

$$x_2' = \min\{2.5, 1.25, 3\} = 1.25 \ (= \theta_2)$$

因为 x_1 是取其上界值的非基变量, 我们应用代换 $x_1 = 4 - x_1'$, 得到

基	x_1'	x_2'	x_3	x_4	x_5	解
z	0	-1	$\frac{5}{2}$	0	$\frac{3}{2}$	$\frac{109}{2}$
x_4	0	1	$\frac{1}{2}$	1	$-\frac{1}{2}$	$\frac{5}{2}$
x_1'	-1	-2	$\frac{3}{2}$	0	$\frac{1}{2}$	$-\frac{5}{2}$

于是, 进基变量 x_2' 成为基变量, 离基变量 x_1 变为取其上界值的非基变量, 由此产生如下表格:

基	x_1'	x_2'	x_3	x_4	x_5	解
z	$\frac{1}{2}$	0	$\frac{7}{4}$	0	$\frac{5}{4}$	$\frac{223}{4}$
x_4	$-\frac{1}{2}$	0	$\frac{5}{4}$	1	$-\frac{1}{4}$	$\frac{5}{4}$
x_2'	$\frac{1}{2}$	1	$-\frac{3}{4}$	0	$-\frac{1}{4}$	$\frac{5}{4}$

最后这张表是可行的且是最优的. 注意, 最后两步能够相互交换, 这意味着我们能够先让 x_2' 成为基变量, 然后再作代换 $x_1 = 4 - x_1'$(请试着做!). 然而, 这里提供的顺序有较少的计算量.

x_1, x_2, x_3 的最优值通过逆代换得到, 分别为 $x_1 = u_1 - x_1' = 4 - 0 = 4$, $x_2 = u_2 - x_2' = 3 - \frac{5}{4} = \frac{7}{4}$, $x_3 = 0$. 最后, 我们得到 $y = l_2 + x_2 = 7 + \frac{7}{4} = \frac{35}{4}$. 相应的目标函数的最优值 z 是 $\frac{223}{4}$.

习题 13.3A

1. 考虑下面的线性规划:

$$\begin{aligned} \max \quad & z = 2x_1 + x_2 \\ \text{s.t.} \quad & x_1 + x_2 \leqslant 3 \\ & 0 \leqslant x_1 \leqslant 2, \ 0 \leqslant x_2 \leqslant 2 \end{aligned}$$

(a) 用图解法求解问题, 并给出在求最优解过程中所得到的极点序列. (可以用 TORA.)

(b) 用上有界算法求解问题, 并说明该方法产生极点序列与图解法产生的极点序列相同 (可以用 TORA 产生相应的迭代).

(c) 上有界算法如何识别极点?

*2. 用有界算法求解下面的问题:

$$\begin{aligned}\max\ & z = 6x_1 + 2x_2 + 8x_3 + 4x_4 + 2x_5 + 10x_6 \\ \text{s.t.}\ & 8x_1 + x_2 + 8x_3 + 2x_4 + 2x_5 + 4x_6 \leqslant 13 \\ & 0 \leqslant x_j \leqslant 1, j = 1, 2, \cdots, 6\end{aligned}$$

3. 用有界算法求解下列问题:

(a)
$$\begin{aligned}\min\ & z = 6x_1 - 2x_2 - 3x_3 \\ \text{s.t.}\ & 2x_1 + 4x_2 + 2x_3 \leqslant 8 \\ & x_1 - 2x_2 + 3x_3 \leqslant 7 \\ & 0 \leqslant x_1 \leqslant 2,\ 0 \leqslant x_2 \leqslant 2,\ 0 \leqslant x_3 \leqslant 1\end{aligned}$$

(b)
$$\begin{aligned}\max\ & z = 3x_1 + 5x_2 + 2x_3 \\ \text{s.t.}\ & x_1 + 2x_2 + 2x_3 \leqslant 10 \\ & 2x_1 + 4x_2 + 3x_3 \leqslant 15 \\ & 0 \leqslant x_1 \leqslant 4,\ 0 \leqslant x_2 \leqslant 3,\ 0 \leqslant x_3 \leqslant 3\end{aligned}$$

4. 在下列问题中, 有一些变量有正的下界. 用有界算法求解下列问题.

(a)
$$\begin{aligned}\max\ & z = 3x_1 + 2x_2 - 2x_3 \\ \text{s.t.}\ & 2x_1 + x_2 + x_3 \leqslant 8 \\ & x_1 + 2x_2 - x_3 \geqslant 3 \\ & 1 \leqslant x_1 \leqslant 3,\ 0 \leqslant x_2 \leqslant 3,\ 2 \leqslant x_3\end{aligned}$$

(b)
$$\begin{aligned}\max\ & z = x_1 + 2x_2 \\ \text{s.t.}\ & -x_1 + 2x_2 \geqslant 0 \\ & 3x_1 + 2x_2 \leqslant 10 \\ & -x_1 + x_2 \leqslant 1 \\ & 1 \leqslant x_1 \leqslant 3,\ 0 \leqslant x_2 \leqslant 1\end{aligned}$$

(c)
$$\begin{aligned}\max\ & z = 4x_1 + 2x_2 + 6x_3 \\ \text{s.t.}\ & 4x_1 - x_2 \leqslant 9 \\ & -x_1 + x_2 + 2x_3 \leqslant 8 \\ & -3x_1 + x_2 + 4x_3 \leqslant 12 \\ & 1 \leqslant x_1 \leqslant 3,\ 0 \leqslant x_2 \leqslant 5,\ 0 \leqslant x_3 \leqslant 2\end{aligned}$$

5. 考虑有界变量问题的矩阵定义. 假定向量 $\boldsymbol{X}$ 被划分成 $(\boldsymbol{X}_z, \boldsymbol{X}_u)$, 其中 $\boldsymbol{X}_u$ 表示在算法进行过程中会在上界处被替换的基变量和非基变量. 因此, 问题可以写成

$$\begin{pmatrix}1 & -\boldsymbol{C}_z & -\boldsymbol{C}_u \\ 0 & \boldsymbol{D}_z & \boldsymbol{D}_u\end{pmatrix}\begin{pmatrix}z \\ \boldsymbol{X}_z \\ \boldsymbol{X}_u\end{pmatrix} = \begin{pmatrix}0 \\ \boldsymbol{b}\end{pmatrix}$$

作变换 $\boldsymbol{X}_u = \boldsymbol{U}_u - \boldsymbol{X}'_u$, 其中 $\boldsymbol{U}_u$ 是 $\boldsymbol{U}$ 的子集, 表示 $\boldsymbol{X}_u$ 的上界, 令 $\boldsymbol{B}$(和 $\boldsymbol{X}_B$) 表示在 $\boldsymbol{X}_u$ 被代换后当前单纯形迭代的基 (和基变量). 证明: 与有界问题相对应的单纯形表具有如下形式:

基	X_z^T	$X_u'^T$	解
z	$C_BB^{-1}D_z - C_z$	$-C_BB^{-1}D_u + C_u$	$C_uB^{-1}b' + C_uU_u$
X_B	$B^{-1}D_z$	$-B^{-1}D_u$	$B^{-1}b'$

其中 $b' = b - D_uU_u$.

6. 在例 13.3-1 中, 完成如下工作.
   (a) 在迭代 1 中, 使用矩阵运算, 验证 $X_B = (x_4, x_1)^T = (\frac{5}{2}, \frac{3}{2})^T$.
   (b) 在迭代 2 中, 说明 B^{-1} 如何由问题的原始数据来计算. 然后用矩阵运算验证基变量 x_4 和 x_2' 的给出的值.
7. 对于上有界变量, 用修正单纯形法 (矩阵形式) 来求解第 3 题的 (a).
8. *有界对偶单纯形算法*. 修正 (4.4.1 节的) 对偶单纯形算法以适应有界变量问题, 其形式如下. 对于所有的 j, 给出其上界约束 $x_j \leqslant u_j$(如果 u_j 是无穷, 则用一个充分大的上界 M 来代替), 将线性规划问题转化成对偶可行 (即原始问题最优) 的形式, 必要时可借助于代换 $x_j = u_j - x_j'$.

   第 1 步 如果当前基变量中任何一个 $(X_B)_i$ 超过它的上界, 作代换 $(X_B)_i = (U_B)_i - (X_B)_i'$. 转到第 2 步.

   第 2 步 如果所有的基变量是可行的, 则停止计算. 否则, 选择离基变量 x_r, 它在基变量中有最负的值. 转到第 3 步.

   第 3 步 用常规的对偶单纯形法 (4.4.1 节) 的最优性条件选择进基变量. 转第 4 步.

   第 4 步 完成基的改变. 转到第 1 步.

   针对下列问题应用给出的算法进行计算:

   (a)
   $$\begin{aligned} \min\ & z = -3x_1 - 2x_2 + 2x_3 \\ \text{s.t.}\ & 2x_1 + x_2 + x_3 \leqslant 8 \\ & -x_1 + 2x_2 + x_3 \geqslant 13 \\ & 0 \leqslant x_1 \leqslant 2, 0 \leqslant x_2 \leqslant 3,\ 0 \leqslant x_3 \leqslant 1 \end{aligned}$$

   (b)
   $$\begin{aligned} \max\ & z = x_1 + 5x_2 - 2x_3 \\ \text{s.t.}\ & 4x_1 + 2x_2 + 2x_3 \leqslant 26 \\ & x_1 + 3x_2 + 4x_3 \geqslant 17 \\ & 0 \leqslant x_1 \leqslant 2,\ 0 \leqslant x_2 \leqslant 3,\ x_3 \geqslant 0 \end{aligned}$$

13.4 对 偶

第 4 章中已处理过对偶问题. 本节介绍有关对偶问题更加精确的处理方法, 并使我们能够检验构成第 4 章后最优分析基础的原始-对偶关系. 这里的介绍还奠定了参数规划的基础.

13.4.1 对偶问题的矩阵定义

假定具有 m 个约束和 n 个变量的等式形式的原始问题定义如下:

$$\max \ z = \boldsymbol{CX}$$
$$\text{s.t.} \ \ \boldsymbol{AX} = \boldsymbol{b}$$
$$\boldsymbol{X} \geqslant \boldsymbol{0}$$

令向量 $\boldsymbol{Y} = (y_1, y_2, \cdots, y_m)$ 表示对偶向量, 按照表 4.2 所述规则, 构造其对偶问题如下:

$$\min \ w = \boldsymbol{Yb}$$
$$\text{s.t.} \ \ \boldsymbol{YA} \geqslant \boldsymbol{C}$$
$$\boldsymbol{Y} \text{ 无限制}$$

$\boldsymbol{YA} \geqslant \boldsymbol{C}$ 中的某些约束可能会覆盖无符号限制的 $\boldsymbol{Y}$.

习题 13.4A

1. 证明: 对偶问题的对偶是原始问题.

*2. 如果原始问题由 $\min \ z = \{\boldsymbol{CX} \mid \boldsymbol{AX} \geqslant \boldsymbol{b}, \ \boldsymbol{X} \geqslant \boldsymbol{0}\}$ 的形式给出, 确定相应的对偶问题.

13.4.2 最优对偶解

本节建立原始问题与对偶问题之间的关系, 并说明如何从原始问题的最优解确定对偶问题的最优解. 令 $\boldsymbol{B}$ 是原始问题的当前最优基, 定义 $\boldsymbol{C}_B$ 是与最优向量 $\boldsymbol{X}_B$ 相对应的目标函数的系数.

定理 13.4-1(弱对偶定理) *对于任意的原始问题与对偶问题的可行解对 $(\boldsymbol{X}, \boldsymbol{Y})$, 极小化问题的目标函数值是极大化问题目标函数的一个上界. 对于最优解对 $(\boldsymbol{X}^*, \boldsymbol{Y}^*)$, 则目标函数值相等.*

证明 可行解对 $(\boldsymbol{X}, \boldsymbol{Y})$ 满足两个问题的所有限制条件. 在极大化问题的约束两边同时左乘 (无限制的)$\boldsymbol{Y}$, 我们得到

$$\boldsymbol{YAX} = \boldsymbol{Yb} = w \tag{1}$$

同样, 对于极小化问题, 不等式约束的两边同时右乘 $\boldsymbol{X}(\geqslant \boldsymbol{0})$, 我们得到

$$\boldsymbol{YAX} \geqslant \boldsymbol{CX}$$

或者

$$\boldsymbol{YAX} \geqslant \boldsymbol{CX} = z \tag{2}$$

(向量 $\boldsymbol{X}$ 的非负性是保证不等式方向的关键.) 结合 (1) 和 (2), 我们得到 $z \leqslant w$ 对于任意的可行解对 $(\boldsymbol{X}, \boldsymbol{Y})$ 成立.

注意, 这个定理并不依赖于将问题标记为原始问题或对偶问题. 重要的是每个问题中最优解的意义. 特别是, 对于任意的可行解对, 极大化问题的目标值不会超过极小化问题的目标值.

该定理的含意是, 由于对任意的可行解有 $z \leqslant w$, 因此, 当两个目标值相等时, 我们就会得到极大化的 z 和极小化的 w. 这个结果的重要性在于, 任何原始问题的可行解与对偶问题的可行解相对于最优解的 "优度" 可以用 $(w-z)$ 与 $\frac{z+w}{2}$ 的比值来检验. 比值 $\frac{2(w-z)}{z+w}$ 越小, 说明两个解越接近最优解. 所建议的单凭经验的方法并不意味着最优目标值是 $\frac{z+w}{2}$.

如果两个问题的目标值中有一个无界, 会怎么样呢? 回答是, 另一个问题一定是不可行的. 如果不是这种情况, 则两个问题均有可行解, 并且关系式 $z \leqslant w$ 成立 —— 这是不可能的结果, 因为由假设, 或者 $z=+\infty$ 或者 $w=-\infty$.

接下来的问题是: 如果一个问题不可行, 另一个问题一定是无界的吗? 不一定. 下面的反例说明了原始问题和对偶问题均不可行 (用图形来验证!).

原始　max $z=\{x_1+x_2 \mid x_1-x_2 \leqslant -1,\ -x_1+x_2 \leqslant -1,\ x_1,x_2 \geqslant 0\}$

对偶　min $w=\{-y_1-y_2 \mid y_1-y_2 \geqslant 1,\ -y_1+y_2 \geqslant 1,\ y_1,y_2 \geqslant 0\}$

定理 13.4-2　*给定原始问题的最优基 $\boldsymbol{B}$ 和与它相对应的目标系数向量 $\boldsymbol{C}_B$, 则对偶问题的最优解是*

$$\boldsymbol{Y}=\boldsymbol{C}_B\boldsymbol{B}^{-1}$$

证明　由定理 13.4-1, 证明只需验证两点: $\boldsymbol{Y}=\boldsymbol{C}_B\boldsymbol{B}^{-1}$ 是对偶问题的可行解, 且 $z=w$.

$\boldsymbol{Y}=\boldsymbol{C}_B\boldsymbol{B}^{-1}$ 的可行性由原始问题的最优性保证, $z_j-c_j \geqslant 0$ 对一切 j 成立, 即

$$\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{A}-\boldsymbol{C} \geqslant \boldsymbol{0}$$

(见 7.2.1 节) 因此, $\boldsymbol{YA}-\boldsymbol{C} \geqslant \boldsymbol{0}$ 或 $\boldsymbol{YA} \geqslant \boldsymbol{C}$, 这说明 $\boldsymbol{Y}=\boldsymbol{C}_B\boldsymbol{B}^{-1}$ 是对偶问题的一个可行解.

下面证明 $w=z$. 注意到

$$w=\boldsymbol{Yb}=\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{b} \tag{1}$$

类似地, 由于原始问题的解为 $\boldsymbol{X}_B=\boldsymbol{B}^{-1}\boldsymbol{b}$, 我们得到

$$z=\boldsymbol{C}_B\boldsymbol{X}_B=\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{b} \tag{2}$$

由关系式 (1) 和 (2), 可得结论 $z=w$.

对偶向量 $\boldsymbol{Y}=\boldsymbol{C}_B\boldsymbol{B}^{-1}$ 有时被称为**对偶** (dual) 或**影子价格** (shadow price), 其名称是由 4.3.1 节中对偶变量的经济学解释演化而来的.

假定 $\boldsymbol{P}_j$ 是 $\boldsymbol{A}$ 的第 j 列, 由定理 13.4-2, 我们注意到

$$z_j-c_j=\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{P}_j-c_j=\boldsymbol{YP}_j-c_j$$

表示对偶问题约束的左端项与右端项之差. 极大化原始问题是以 $z_j - c_j < 0$ 至少有一个 j 成立开始, 这意味着相应的对偶约束 $\boldsymbol{YP}_j \geqslant c_j$ 不满足. 当原始问题达到最优解时, 我们得到 $z_j - c_j \geqslant 0$ 对所有的 j 成立, 这意味着相应的对偶解 $\boldsymbol{Y} = \boldsymbol{C}_B\boldsymbol{B}^{-1}$ 已成为可行解. 因此, 当原始问题寻找最优性时, 其对偶问题自动地寻找可行性. 这一点是**对偶单纯形法**(4.4.1 节) 的发展基础, 对偶单纯形方法以优于最优但不可行的解开始, 并在随后的迭代中保持最优性, 直到在最后的迭代中获得可行性为止. 这与 (原始) 单纯形法 (第 3 章) 形成明显的对照; 原始单纯形法保持劣于最优但可行的解开始迭代, 直到获得最优性为止.

例 13.4-1

如下线性规划最优基为 $\boldsymbol{B} = (\boldsymbol{P}_1, \boldsymbol{P}_4)$. 写出其对偶问题, 并用原始问题的最优基求出对偶问题的最优解.

$$\begin{aligned}
\max\ \ & z = 3x_1 + 5x_2 \\
\text{s.t.}\ \ & x_1 + 2x_2 + x_3 = 5 \\
& -x_1 + 3x_2 + x_4 = 2 \\
& x_1, x_2, x_3, x_4 \geqslant 0
\end{aligned}$$

对偶问题是

$$\begin{aligned}
\min\ \ & w = 5y_1 + 2y_2 \\
\text{s.t.}\ \ & y_1 - y_2 \geqslant 3 \\
& 2y_1 + 3y_2 \geqslant 5 \\
& y_1, y_2 \geqslant 0
\end{aligned}$$

我们有 $\boldsymbol{X}_B = (x_1, x_4)^{\mathrm{T}}$ 和 $\boldsymbol{C}_B = (3, 0)$. 最优基及其逆分别为

$$\boldsymbol{B} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}, \quad \boldsymbol{B}^{-1} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

与其相对应的原始问题和对偶问题的解分别为

$$(x_1, x_4)^{\mathrm{T}} = \boldsymbol{B}^{-1}\boldsymbol{b} = (5, 7)^{\mathrm{T}}$$
$$(y_1, y_2) = \boldsymbol{C}_B\boldsymbol{B}^{-1} = (3, 0)$$

两个解均是可行的, 并且 $z = w = 15$(请验证!). 因此, 两个解均是最优的.

习题 13.4B

1. 验证: 在定理 13.4-1 后面给出的数值例子的对偶问题是正确的. 然后用图解法验证原始问题和对偶问题均无可行解.

2. 考虑如下线性规划:

$$\begin{aligned} \max\ & z = 50x_1 + 30x_2 + 10x_3 \\ \text{s.t.}\ & 2x_1 + x_2 = 1 \\ & 2x_2 = -5 \\ & 4x_1 + x_3 = 6 \\ & x_1, x_2, x_3 \geqslant 0 \end{aligned}$$

(a) 写出其对偶问题.　　(b) 用观察的方法说明原始问题不可行.

(c) 说明 (a) 中的对偶问题无界.

(d) 从第 1 题和第 2 题, 研究出一个关于原始问题和对偶问题不可行与无界的一般结论.

3. 考虑如下线性规划:

$$\begin{aligned} \max\ & z = 5x_1 + 12x_2 + 4x_3 \\ \text{s.t.}\ & 2x_1 - x_2 + 3x_3 = 2 \\ & x_1 + 2x_2 + x_3 + x_4 = 5 \\ & x_1, x_2, x_3, x_4 \geqslant 0 \end{aligned}$$

(a) 写出其对偶问题.

(b) 在下列每种情况中, 对于原始问题, 首先验证给出的基 $\boldsymbol{B}$ 是可行的. 接下来, 用 $\boldsymbol{Y} = \boldsymbol{C}_B\boldsymbol{B}^{-1}$ 计算相应的对偶变量, 并验证原始问题的解是否为最优解.

(i) $\boldsymbol{B} = (\boldsymbol{P}_4, \boldsymbol{P}_3)$　　(ii) $\boldsymbol{B} = (\boldsymbol{P}_2, \boldsymbol{P}_3)$

(iii) $\boldsymbol{B} = (\boldsymbol{P}_1, \boldsymbol{P}_2)$　　(iv) $\boldsymbol{B} = (\boldsymbol{P}_1, \boldsymbol{P}_4)$

4. 考虑如下线性规划:

$$\begin{aligned} \max\ & z = 2x_1 + 4x_2 + 4x_3 - 3x_4 \\ \text{s.t.}\ & x_1 + x_2 + x_3 = 4 \\ & x_1 + 4x_2 + x_4 = 8 \\ & x_1, x_2, x_3, x_4 \geqslant 0 \end{aligned}$$

(a) 写出其对偶问题.

(b) 对所有的非基向量 $\boldsymbol{P}_j$, 计算 $z_j - c_j$, 验证 $\boldsymbol{B} = (\boldsymbol{P}_2, \boldsymbol{P}_3)$ 是最优基.

(c) 求相应的对偶问题的最优解.

***5.** 一个线性规划模型包括 2 个变量 x_1 和 x_2, 以及 3 个类型为 $\leqslant$ 的不等式约束. 与约束对应的松弛变量为 x_3, x_4, x_5. 假定最优基为 $\boldsymbol{B} = (\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3)$, 且它的逆为

$$\boldsymbol{B}^{-1} = \begin{pmatrix} 0 & -1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & -1 \end{pmatrix}$$

原始问题和对偶问题的最优解是

$$\boldsymbol{X}_B = (x_1, x_2, x_3)^{\mathrm{T}} = (2, 6, 2)^{\mathrm{T}}$$

$$\boldsymbol{Y} = (y_1, y_2, y_3) = (0, 3, 2)$$

用原始问题和对偶问题两种方法确定目标函数的最优值.

6. 对于原始问题和对偶问题的最优解, 证明下面的关系式:

$$\sum_{i=1}^{m} c_i\left(\boldsymbol{B}^{-1}\boldsymbol{P}_k\right)_i = \sum_{i=1}^{m} y_i a_{ik}$$

其中 $\boldsymbol{C}_B = (c_1, c_2, \cdots, c_m)$, $\boldsymbol{P}_k = (a_{1k}, a_{2k}, \cdots, a_{mk})^{\mathrm{T}}$, $k = 1, 2, \cdots, n$, 且 $\left(\boldsymbol{B}^{-1}\boldsymbol{P}_k\right)_i$ 是 $\boldsymbol{B}^{-1}\boldsymbol{P}_k$ 的第 i 个元素.

*7. 写出

$$\max\ z = \{\boldsymbol{CX} \mid \boldsymbol{AX} = \boldsymbol{b}, \boldsymbol{X}\text{无限制}\}$$

的对偶问题.

8. 证明:

$$\max\ z = \{\boldsymbol{CX} \mid \boldsymbol{AX} \leqslant \boldsymbol{b}, \boldsymbol{0} < \boldsymbol{L} \leqslant \boldsymbol{X} \leqslant \boldsymbol{U}\}$$

的对偶问题总有可行解.

13.5　参数线性规划

参数线性规划是 4.5 节提出的后最优分析的延伸. 它详细研究目标函数系数和约束右端项在预定的连续变化区域内对于最优解的影响效果.

令 $\boldsymbol{X} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$ 并定义线性规划如下:

$$\max\ z = \left\{\boldsymbol{CX} \mid \sum_{j=1}^{n} \boldsymbol{P}_j x_j = \boldsymbol{b},\ \boldsymbol{X} \geqslant \boldsymbol{0}\right\}$$

在参数分析中, 用参数化函数 $\boldsymbol{C}(t)$ 和 $\boldsymbol{b}(t)$ 分别表示目标函数 $\boldsymbol{C}$ 和右端项向量 $\boldsymbol{b}$, 其中 t 是可变化的参数. 从数学角度上讲, 可以假定 t 为任意的正值或负值. 然而, 在实际中, t 通常表示时间, 因此, 它是非负的. 在这里的介绍中, 我们将假定 $t \geqslant 0$.

参数分析的一般思想是以 $t = 0$ 时的最优解开始. 然后, 我们用单纯形法的最优性和可行性条件确定这样一个区域 $0 \leqslant t \leqslant t_1$: 在这个区域中, $t = 0$ 时的解仍保持最优与可行. 在这种情况下, 称 t_1 为**临界值** (critical value). 这个过程由确定一系列临界值及其相应的最优可行解来延续; 它将终止于 $t = t_r$, 如果它显示出当 $t > t_r$ 时最后的最优解保持不变, 或者超出这个临界值时可行解不存在.

13.5.1　C 中的参数变化

令 $\boldsymbol{X}_{B_i}$, $\boldsymbol{B}_i$, $\boldsymbol{C}_{B_i}(t)$ 是确定对应于临界值 t_i 的最优解的基本要素 (计算以 $t_0 = 0$ 且最优基为 $\boldsymbol{B}_0$ 开始). 下一步, 确定临界值 t_{i+1} 和相应的最优基 (只要有一个存在). 因为 $\boldsymbol{C}$ 的变化只影响问题的最优性条件, 因此, 对于 $t \geqslant t_i$, 当前解 $\boldsymbol{X}_{B_i} = \boldsymbol{B}_i^{-1}\boldsymbol{b}$ 仍保持最优, 只要简约费用 $z_j(t) - c_j(t)$ 满足下面的最优性条件:

$$z_j(t) - c_j(t) = \boldsymbol{C}_{B_i}(t)\boldsymbol{B}_i^{-1}\boldsymbol{P}_j - c_j(t) \geqslant 0,\ \forall j$$

t_{i+1} 的值等于满足所有最优性条件中最大的 $t > t_i$ 的值.

注意, 不等式并不需要 $C(t)$ 是关于 t 的线性函数. 任何 $C(t)$ 函数 (线性或非线性) 均是可接受的. 然而, 对于非线性函数, 由不等式得到的数值运算可能是繁琐的 (见习题 13.5A 的第 5 题, 它足以说明这种情况.)

例 13.5-1

$$\begin{aligned} \max\ & z=(3-6t)x_1+(2-2t)x_2+(5+5t)x_3 \\ \text{s.t.}\ & x_1+2x_2+x_3 \leqslant 40 \\ & 3x_1+2x_3 \leqslant 60 \\ & x_1+4x_2 \leqslant 30 \\ & x_1,x_2,x_3 \geqslant 0 \end{aligned}$$

我们有

$$C(t)=(3-6t,2-2t,5+5t),\quad t\geqslant 0$$

用变量 x_4, x_5, x_6 作为这 3 个约束的松弛变量.

$t=t_0=0$ 时的最优解

基	x_1	x_2	x_3	x_4	x_5	x_6	解
z	4	0	0	1	2	0	160
x_2	$-\frac{1}{4}$	1	0	$\frac{1}{2}$	$-\frac{1}{4}$	0	5
x_3	$\frac{3}{2}$	0	1	0	$\frac{1}{2}$	0	30
x_6	2	0	0	-2	1	1	10

$$X_{B_0}=(x_2,x_3,x_6)^{\mathrm{T}}=(5,30,10)^{\mathrm{T}}$$

$$C_{B_0}(t)=(2-2t,5+5t,0)$$

$$B_0^{-1}=\begin{pmatrix} \frac{1}{2} & -\frac{1}{4} & 0 \\ 0 & \frac{1}{2} & 0 \\ -2 & 1 & 1 \end{pmatrix}$$

对于当前非基向量 P_1, P_4, P_5, 最优性条件是

$$\{C_{B_0}(t)B_0^{-1}P_j-c_j(t)\}_{j=1,4,5}=(4+14t,1-t,2+3t)\geqslant \mathbf{0}$$

因此, X_{B_0} 保持最优仅当下列条件满足:

$$\begin{aligned} 4+14t &\geqslant 0 \\ 1-t &\geqslant 0 \\ 2+3t &\geqslant 0 \end{aligned}$$

因为 $t \geqslant 0$, 第二个不等式给出 $t \leqslant 1$, 并对于所有 $t \geqslant 0$ 其余两个不等式均得到满足. 因此, 我们得到 $t_1 = 1$, 这意味着, 对于 $0 \leqslant t \leqslant 1$, $\boldsymbol{X}_{B_0}$ 保持最优 (且可行).

当 $t = 1$ 时, 简约费用 $z_4(t) - c_4(t) = 1 - t$ 等于 0; 对于 $t > 1$, 简约费用为负数. 因此, 对于 $t > 1$, $\boldsymbol{P}_4$ 一定进基. 在这种情况下, $\boldsymbol{P}_2$ 一定离基 (见 $t = 0$ 时的最优单纯形表). 通过让 $\boldsymbol{P}_4$ 进基, 在 $t = 1$ 处得到另一个新的基本解 $\boldsymbol{X}_{B_1}$, 即 $\boldsymbol{X}_{B_1} = (x_4, x_3, x_6)^{\mathrm{T}}$ 和 $\boldsymbol{B}_1 = (\boldsymbol{P}_4, \boldsymbol{P}_3, \boldsymbol{P}_6)$.

$t = t_1 = 1$ 时的另一个最优基

$$\boldsymbol{B}_1 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{B}_1^{-1} = \begin{pmatrix} 1 & -\frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

因此,

$$\boldsymbol{X}_{B_1} = (x_4, x_3, x_6)^{\mathrm{T}} = \boldsymbol{B}_1^{-1}\boldsymbol{b} = (10, 30, 30)^{\mathrm{T}}$$

$$\boldsymbol{C}_{B_1}(t) = (0, 5 + 5t, 0)$$

相应的非基向量是 $\boldsymbol{P}_1$, $\boldsymbol{P}_2$, $\boldsymbol{P}_5$, 因此我们有

$$\left\{\boldsymbol{C}_{B_1}(t)\boldsymbol{B}_1^{-1}\boldsymbol{P}_j - c_j(t)\right\}_{j=1,2,5} = \left(\frac{9+27t}{2}, -2+2t, \frac{5+5t}{2}\right) \geqslant \boldsymbol{0}$$

按照这些条件, 对于所有的 $t \geqslant 1$, 基本解 $\boldsymbol{X}_{B_1}$ 仍然是最优的. 观察到, 最优性条件 $-2+2t \geqslant 0$ 自动地 "记住" $\boldsymbol{X}_{B_1}$ 在 t 的以最后的临界值 $t_1 = 1$ 开始的区域内是最优的. 这是在参数规划的计算中总能出现的一种情况.

对于 t 在整个区域的情形, 其最优解总结如下. 由直接替换计算出 z 的值.

t	x_1	x_2	x_3	z
$0 \leqslant t \leqslant 1$	0	5	30	$160 + 140t$
$t \geqslant 1$	0	0	30	$150 + 150t$

习题 13.5A

*1. 在例 13.5-1 中, 假设 t 无符号限制. 确定 t 的区间, 使得 $\boldsymbol{X}_{B_0}$ 保持最优.

2. 求解例 13.5-1, 假定目标函数已知如下:

*(a) $\max\ z = (3 + 3t)x_1 + 2x_2 + (5 - 6t)x_3$

(b) $\max\ z = (3 - 2t)x_1 + (2 + t)x_2 + (5 + 2t)x_3$

(c) $\max\ z = (3 + t)x_1 + (2 + 2t)x_2 + (5 - t)x_3$

3. 研究如下参数线性规划最优解的变化, 且 $t \geqslant 0$.

$$\begin{aligned} \min \ & z = (4-t)x_1 + (1-3t)x_2 + (2-2t)x_3 \\ \text{s.t.} \ & 3x_1 + x_2 + 2x_3 = 3 \\ & 4x_1 + 3x_2 + 2x_3 \geqslant 6 \\ & x_1 + 2x_2 + 5x_3 \leqslant 4 \\ & x_1, x_2, x_3 \geqslant 0 \end{aligned}$$

4. 在本节的分析中, 假定由 (原始) 单纯形法获得 $t=0$ 处线性规划的最优解. 在一些问题中, 由对偶单纯形法 (4.4.1 节) 获得最优解可能更加方便. 说明在这种情况下如何完成参数分析. 然后, 分析例 4.4-1, 假定目标函数按如下形式给出:

$$\min \ z = (3+t)x_1 + (2+4t)x_2 + x_3, \quad t \geqslant 0$$

***5.** 在例 13.5-1 中, 假定目标函数关于 t 是非线性的, 并定义为

$$\max \ z = (3+2t^2)x_1 + (2-2t^2)x_2 + (5-t)x_3$$

确定第一个临界值 t_1.

13.5.2 b 中的参数变化

参数化右端项 $b(t)$ 只能影响到问题的可行性. 因此, t 的临界值由下列条件确定:

$$\boldsymbol{X}_B(t) = \boldsymbol{B}^{-1}\boldsymbol{b}(t) \geqslant \boldsymbol{0}$$

例 13.5-2

$$\begin{aligned} \max \ & z = 3x_1 + 2x_2 + 5x_3 \\ \text{s.t.} \ & x_1 + 2x_2 + x_3 \leqslant 40 - t \\ & 3x_1 \qquad\ + 2x_3 \leqslant 60 + 2t \\ & x_1 + 4x_2 \qquad \leqslant 30 - 7t \\ & x_1, x_2, x_3 \geqslant 0 \end{aligned}$$

假定 $t \geqslant 0$.

当 $t = t_0 = 0$ 时, 问题与例 13.5-1 相同. 因此我们有

$$\boldsymbol{X}_{B_0} = (x_2, x_3, x_6)^{\mathrm{T}} = (5, 30, 10)^{\mathrm{T}}$$

$$\boldsymbol{B}_0^{-1} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{4} & 0 \\ 0 & \frac{1}{2} & 0 \\ -2 & 1 & 1 \end{pmatrix}$$

为了确定第一个临界值 t_1, 应用可行性条件 $\boldsymbol{X}_{B_0}(t) = \boldsymbol{B}_0^{-1}\boldsymbol{b}(t) \geqslant 0$, 得到

$$\begin{pmatrix} x_2 \\ x_3 \\ x_6 \end{pmatrix} = \begin{pmatrix} 5-t \\ 30+t \\ 10-3t \end{pmatrix} \geqslant \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

对于 $t \leqslant \frac{10}{3}$, 这些不等式成立, 这意味着 $t_1 = \frac{10}{3}$, 且对于区间 $0 \leqslant t \leqslant \frac{10}{3}$, 基 $\boldsymbol{B}_0$ 保持可行. 然而, 基变量 x_2, x_3, x_6 的值将像前面一样随 t 变化.

基变量 $x_6(= 10 - 3t)$ 的值将在 $t = t_1 = \frac{10}{3}$ 处等于 0, 且对于 $t > \frac{10}{3}$, 它将变为负值. 因此, 在 $t = \frac{10}{3}$ 处, 我们能够应用修正对偶单纯形法 (其细节请见习题 13.2B 的第 5 题) 确定另一个基 $\boldsymbol{B}_1$. x_6 是离基变量.

$t = t_1 = \frac{10}{3}$ 时的另一个基

已知 x_6 是离基变量, 确定进基变量如下:

$$\boldsymbol{X}_{B_0} = (x_2, x_3, x_6)^{\mathrm{T}}, \quad \boldsymbol{C}_{B_0} = (2, 5, 0)$$

因此,

$$\{z_j - c_j\}_{j=1,4,5} = \left\{\boldsymbol{C}_{B_0}\boldsymbol{B}_0^{-1}\boldsymbol{P}_j - c_j\right\}_{j=1,4,5} = (4, 1, 2)$$

下一步, 对于非基变量 x_j, $j = 1, 4, 5$, 我们计算

$$\begin{aligned}(\boldsymbol{B}_0^{-1}\text{中与 } x_6 \text{ 对应的行})(\boldsymbol{P}_1, \boldsymbol{P}_4, \boldsymbol{P}_5) &= (\boldsymbol{B}_0^{-1}\text{的第 3 行})(\boldsymbol{P}_1, \boldsymbol{P}_4, \boldsymbol{P}_5)\\ &= (-2, 1, 1)(\boldsymbol{P}_1, \boldsymbol{P}_4, \boldsymbol{P}_5) = (2, -2, 1)\end{aligned}$$

因此, 进基变量相对应的值是

$$\theta = \min\left\{-, \left|\frac{1}{-2}\right|, -\right\} = \frac{1}{2}$$

所以, $\boldsymbol{P}_4$ 是进基向量. 另一个基本解及其 $\boldsymbol{B}_1$ 和 $\boldsymbol{B}_1^{-1}$ 分别是

$$\boldsymbol{X}_{B_1} = (x_2, x_3, x_4)^{\mathrm{T}}$$

$$\boldsymbol{B}_1 = (\boldsymbol{P}_2, \boldsymbol{P}_3, \boldsymbol{P}_4) = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 2 & 0 \\ 4 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{B}_1^{-1} = \begin{pmatrix} 0 & 0 & \frac{1}{4} \\ 0 & \frac{1}{2} & 0 \\ 1 & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix}$$

下一个临界值 t_2 由可行性条件 $\boldsymbol{X}_{B_1}(t) = \boldsymbol{B}_1^{-1}\boldsymbol{b}(t) \geqslant \boldsymbol{0}$ 确定, 于是

$$\begin{pmatrix} x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} \frac{30-7t}{4} \\ 30 + t \\ \frac{-10+3t}{2} \end{pmatrix} \geqslant \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

这些条件说明 $\boldsymbol{B}_1$ 在 $\frac{10}{3} \leqslant t \leqslant \frac{30}{7}$ 内保持可行.

当 $t = t_2 = \frac{30}{7}$ 时, 另一个基可由修正对偶单纯形法获得. x_2 是离基变量, 因为它对应于产生临界值 t_2 的条件.

$t = t_2 = \frac{30}{7}$ 时的另一个基

已知 x_2 是离基变量, 确定进基变量如下:

$$\boldsymbol{X}_{B_1} = (x_2, x_3, x_4)^{\mathrm{T}}, \quad \boldsymbol{C}_{B_1} = (2, 5, 0)$$

因此,

$$\{z_j - c_j\}_{j=1,5,6} = \left\{\boldsymbol{C}_{B_1}\boldsymbol{B}_1^{-1}\boldsymbol{P}_j - c_j\right\}_{j=1,5,6} = \left(5, \frac{5}{2}, \frac{1}{2}\right)$$

下一步, 对于非基变量 x_j, $j = 1,5,6$, 我们计算

$$\begin{aligned}(\boldsymbol{B}_1^{-1}\text{中与 } x_2 \text{ 对应的行})(\boldsymbol{P}_1, \boldsymbol{P}_5, \boldsymbol{P}_6) &= (\boldsymbol{B}_1^{-1}\text{的第一行})(\boldsymbol{P}_1, \boldsymbol{P}_5, \boldsymbol{P}_6) \\ &= \left(0, 0, \tfrac{1}{4}\right)(\boldsymbol{P}_1, \boldsymbol{P}_5, \boldsymbol{P}_6) = \left(\tfrac{1}{4}, 0, \tfrac{1}{4}\right)\end{aligned}$$

因为 $\left(\frac{1}{4}, 0, \frac{1}{4}\right)$ 的所有元素均 $\geqslant 0$, 且对于 $t > \frac{30}{7}$, 问题没有可行解, 所以参数分析终止于 $t = t_2 = \frac{30}{7}$.

最优解概括如下.

t	x_1	x_2	x_3	z
$0 \leqslant t \leqslant \frac{10}{3}$	0	$5 - t$	$30 + t$	$160 + 3t$
$\frac{10}{3} \leqslant t \leqslant \frac{30}{7}$	0	$\frac{30-7t}{4}$	$30 + t$	$165 + \frac{3}{2}t$
$t > \frac{30}{7}$	(没有可行解存在)			

习题 13.5B

*1. 在例 13.5-2 中, 求第一个临界值 t_1, 并针对下列情况确定 $\boldsymbol{B}_1$ 的基向量:

*(a) $\boldsymbol{b}(t) = (40 + 2t, 60 - 3t, 30 + 6t)^{\mathrm{T}}$

(b) $\boldsymbol{b}(t) = (40 - t, 60 + 2t, 30 - 5t)^{\mathrm{T}}$

*2. 研究如下参数线性规划最优解的变化, 已知 $t \geqslant 0$.

$$\begin{aligned}\min\ & z = 4x_1 + x_2 + 2x_3 \\ \text{s.t. } & 3x_1 + x_2 + 2x_3 = 3 + 3t \\ & 4x_1 + 3x_2 + 2x_3 \geqslant 6 + 2t \\ & x_1 + 2x_2 + 5x_3 \leqslant 4 - t \\ & x_1, x_2, x_3 \geqslant 0\end{aligned}$$

3. 在本节的分析中, 假定 $t = 0$ 处线性规划的最优解是通过 (原始) 单纯形法获得的. 在一些问题中, 由对偶单纯形方法 (4.4.1 节) 获得最优解可能更加方便. 说明在这种情况下如何完成参数分析. 然后, 分析例 4.4-1, 假定 $t \geqslant 0$ 且右端项向量是

$$\boldsymbol{b}(t) = (3 + 2t, 6 - t, 3 - 4t)^{\mathrm{T}}$$

4. 求解第 2 题, 假定右端项变为

$$\boldsymbol{b}(t) = (3 + 3t^2, 6 + 2t^2, 4 - t^2)^{\mathrm{T}}$$

进一步假定, t 可以是正、零或负.

参 考 文 献

Bazaraa, M., J. Jarvis, and H. Sherali, *Linear Programming and Network Flows*, 2nd ed., Wiley, New York, 1990.

Chvàtal, V., *Linear Programming*, Freeman, San Francisco, 1983.

Nering, E., and A. Tucker, *Linear Programming and Related Problems*, Academic Press, Boston, 1992.

Saigal, R., *Linear Programming: A Modern Integrated Analysis*, Kluwer Academic Publishers, Boston, 1995.

Vanderbei, R., *Linear Programming: Foundation and Extensions*, 2nd ed, Kluwer Academic Publishers, Boston, 2001.

第 14 章　概率论基础复习

本章导读　本章复习概率原理、随机变量和概率分布. 已经学习过基础概率统计课程的读者, 可以跳过这一章. 此外, 本章还简要介绍了本书经常用到的 5 种概率分布: 二项分布、泊松分布、均匀分布、指数分布和正态分布. 我们还开发了一个基于电子表格的统计表 (文件 StatTables.xls), 用来自动计算 16 不同分布的均值、标准差、概率和百分位数. 本章还提供了另一个电子表格, 用来对实验数据画出直方图 (文件 excelMeanVar.xls).

本章包括 12 个例题、2 个电子表格和 44 个节后习题. AMPL/Excel/Solver/TORA 程序放在文件夹 ch12Files 下.

14.1　概 率 原 理

概率用来分析一项**实验** (experiment) 产生的随机结果. 所有这些可能结果的总体称为**样本空间** (sample space), 样本空间的一个子集称为一个**事件** (event). 例如, 掷一次 (6 面) 骰子的可能结果有 1, 2, 3, 4, 5, 6 点. 集合 $\{1,2,3,4,5,6\}$ 就定义了相应的样本空间, 掷一次骰子出现偶数值 (2, 4, 6) 就是一个事件.

实验还可能产生连续的样本空间. 例如, 电器元件的正常连续运行时间可以是任一非负值.

如果事件 E 在 n 次实验中发生了 m 次, 则事件 E 发生的概率 $P\{E\}$ 定义为

$$P\{E\} = \lim_{n\to\infty} \frac{m}{n}$$

这一定义说明了, 假如这个实验重复*无限次* ($n \to \infty$), 则所需要的概率可以用 $\frac{m}{n}$ 来表示. 可以通过投掷硬币并观察它的结果 [正面 (H) 或反面 (T)] 来验证这一定义. 重复这一实验的时间越长, $P\{H\}$(或 $P\{T\}$) 的估计值越接近于理论值 0.5.

按照定义,

$$0 \leqslant P\{E\} \leqslant 1$$

一个事件称为不可能事件, 如果 $P\{E\} = 0$. 若 $P\{E\} = 1$, 则称其为必然事件. 例如, 在一次掷 6 面骰子的实验中, 掷出 7 点是不可能事件, 而掷出 1 到 6 的某个整数值则是必然事件.

习题 14.1A

*1. 为了了解高三年级数学成绩与选修工程学院课程之间的相互关系, 人们对阿肯色州的一些高中开展了一项调查. 在被调查的 1 000 名高三年级学生中, 有 400 名学生学完了数学课程, 而选修工程课程的情况表明, 在 1 000 名高三学生中, 有 150 名学生学习过数学, 有 29 名没学过. 求下列事件的概率:
 (a) 一名学过数学的学生选修了工程课程. 该学生没有选修工程课程.
 (b) 一名学生既没学过数学, 也没选修工程课程.
 (c) 一名学生没有学工程课程.

*2. 考虑随机召集 n 个人, 求出最小的 n, 使得很有可能有两个或两个以上的人生日相同 (提示: 假定不跨越年份, 且一年中每一天都等可能是某个人的生日).

*3. 假定有两个或两个以上的人和你的生日相同, 回答第 2 题.

14.1.1 概率的加法律

对两个事件 E 和 F, $E+F$(或 $E\cup F$) 表示 E 和 F 的**并** (union), EF (或 $E\cap F$) 表示它们的**交** (intersection). **事件 E 和 F 是互斥的** (mutually exclusive), 如果它们之间没有交, 即一个事件的发生排除了另一个事件的发生. 根据上述定义, 概率的加法律可以表述为

$$P\{E+F\}=\begin{cases} P\{E\}+P\{F\}, & E\text{ 和 }F\text{ 互斥} \\ P\{E\}+P\{F\}-P\{EF\}, & \text{否则} \end{cases}$$

$P\{EF\}$ 是事件 E 和 F 同时发生的概率.

例 14.1-1

考虑掷骰子实验. 实验的样本空间是 $\{1,2,3,4,5,6\}$. 对于一只均匀的骰子, 我们有

$$P\{1\}=P\{2\}=P\{3\}=P\{4\}=P\{5\}=P\{6\}=\frac{1}{6}$$

定义

$$E=\{1,2,3,4\}$$
$$F=\{3,4,5\}$$

E 和 F 都出现了 3 和 4. 因此, $EF=\{3,4\}$, 这样

$$P\{E\}=P\{1\}+P\{2\}+P\{3\}+P\{4\}=\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=\frac{2}{3}$$
$$P\{F\}=P\{3\}+P\{4\}+P\{5\}=\frac{1}{2}$$
$$P\{EF\}=P\{3\}+P\{4\}=\frac{1}{3}$$

由此得出

$$P\{E+F\}=P\{E\}+P\{F\}-P\{EF\}=\frac{2}{3}+\frac{1}{2}-\frac{1}{3}=\frac{5}{6}$$

直观上这个结果是对的, 因为 $\{E+F\}=\{1,2,3,4,5\}$, 它发生的概率就是 $\frac{5}{6}$.

习题 14.1B

1. 把一只均匀的 6 面骰子掷两次. 令 E 和 F 代表两次投掷的结果, 计算下列事件的概率.
   (a) E 和 F 的和为 11. (b) E 和 F 的和是偶数.
   (c) E 和 F 的和是奇数且大于 3. (d) E 是小于 6 的偶数, F 是大于 1 的奇数.
   (e) E 大于 2 且 F 小于 4. (f) E 等于 4, E 和 F 的和是奇数.
2. 假定你独立地掷两次骰子并记录下每次朝上的点数, 求下列概率.
   (a) 两个数都是偶数的概率. (b) 两数之和为 10 的概率.
   (c) 两数之差至少为 3 的概率.

*3. 将一枚均匀的硬币朝上抛掷 7 次. 如果 1 次正面朝上之前连续 3 次反面, 你就会赢 $100, 那么你赢钱的机会是多少?

*4. Ann(女), Jim(男), John(男) 和 Liz(女) 要进行一次壁球循环赛. Ann 有双倍的可能性打败 Jim, Jim 和 John 处在同一水平上, 而 Liz 过去打赢 John 的纪录是每三次赢一次. 求下列概率.
   (a) Jim 赢得这次循环赛的概率. (b) 女选手赢得这次循环赛的概率.
   (c) 没有女选手获胜的概率.

14.1.2 条件概率定律

给定两个事件 E 和 F, $\mathrm{P}\{F\} > 0$, 已知 F 发生时事件 E 的条件概率 $\mathrm{P}\{E|F\}$ 被定义为

$$\mathrm{P}\{E|F\} = \frac{\mathrm{P}\{EF\}}{\mathrm{P}\{F\}}, \quad \mathrm{P}\{F\} > 0$$

假如 E 是 F 的子集 (E 包括在 F 中), 则 $\mathrm{P}\{EF\} = \mathrm{P}\{E\}$.

两个事件 E 和 F 是*独立的*, 当且仅当

$$\mathrm{P}\{E|F\} = \mathrm{P}\{E\}$$

在这种情况下, 条件概率定律简化成

$$\mathrm{P}\{EF\} = \mathrm{P}\{E\}\mathrm{P}\{F\}$$

例 14.1-2

你和另一个人玩一个二人游戏, 其中另一个人掷骰子, 你看不到骰子, 但对方会告诉你有关结果的信息. 你的任务是预测出每次投掷的点数. 假设告诉你朝上的是一个偶数, 求结果是 6 点的概率.

令 $E = \{6\}$, 并定义 $F = \{2, 4, 6\}$, 因此

$$\mathrm{P}\{E|F\} = \frac{\mathrm{P}\{EF\}}{\mathrm{P}\{F\}} = \frac{\mathrm{P}\{E\}}{\mathrm{P}\{F\}} = \left(\frac{1/6}{1/2}\right) = \frac{1}{3}$$

注意到 $\mathrm{P}\{EF\} = \mathrm{P}\{E\}$, 因为 E 是 F 的子集.

习题 14.1C

1. 在例 14.1-2 中, 假设对方告诉你结果点数小于 6.
   (a) 求得到偶数点的概率.　　(b) 求得到大于 1 的奇数点的概率.
2. 沃尔玛公司的股票在纽约证券市场上用 WMS 编号进行交易. 历史上, WMS 的股价 60% 的时候随着道琼斯指数的上升而上涨, 有 25% 的时候随道琼斯指数的下降而下跌, 还有 5% 的机会当道琼斯指数下降但 WMS 股价上扬, 以及 10% 的时候道琼斯指数上升而 WMS 反跌.
   (a) 求不论道琼斯指数如何而 WMS 股价上涨的概率.
   (b) 求已知道琼斯指数上升时 WMS 股价上涨的概率.
   (c) 求已知道琼斯指数下降时 WMS 股价下跌的概率.

*3. 获得 26 分或以上 ACT 成绩的高中毕业班学生可报考 A 和 B 两所大学. 被 A 大学录取的概率是 0.4, 被 B 大学录取的概率为 0.25, 被这两所大学都接受的机会只有 15%.
   (a) 已知某学生已经被 A 大学录取了, 求该学生被 B 大学接受的概率.
   (b) 已知该学生被 B 大学接受了, A 大学录取他的概率是多少?

4. 证明假如概率 $\mathrm{P}\{A|B\} = \mathrm{P}\{A\}$, 则 A 和 B 必为独立的.
5. 贝叶斯定理 (Bayes' theorem)[①]　给定两个事件 A 和 B, 证明

$$\mathrm{P}\{A|B\} = \frac{\mathrm{P}\{B|A\}\mathrm{P}\{A\}}{\mathrm{P}\{B\}}, \mathrm{P}\{B\} > 0$$

6. 一家零售商销售的电池有 75% 来自工厂 A, 25% 来自工厂 B. A 和 B 生产的不合格产品的比率分别是 1% 和 2%. 一位顾客刚刚随机地从该零售商处买到一块电池.
   (a) 该电池是不合格品的概率有多少?
   (b) 假如你购买的电池是不合格品, 它来自工厂 A 的概率是多少? (提示: 利用第 5 题的贝叶斯定理)

*7. 统计表明, 有 70% 的男人都患有某种形式的前列腺癌. 对有病症的人做 PSA 检查时有 90% 的时候呈阳性, 对健康人检查时有 10% 的时候呈阳性. 一个人检查呈阳性且确实患前列腺癌的概率是多少?

14.2　随机变量与概率分布

一次实验的结果既可以是数值, 也可以是某种数字编码. 例如: 掷骰子的结果是 1 到 6 的自然数; 而对一件物品的检验产生好的和坏的两个结果. 此时可以用数字代码 (0, 1) 来表示 (坏, 好). 这些结果的数字表示产生了所谓的**随机变量** (random variable).

随机变量 x 可以是**离散的** (discrete), 也可以是**连续的** (continuous). 例如, 与掷骰子有关的随机变量是离散的, 其中 $x = 1, 2, 3, 4, 5, 6$; 而一个服务设施的到达间隔时间则是连续的, 其中 $x \geqslant 0$.

---

① 11.2.2 节给出了贝叶斯定理的详细表述.

每个连续的或离散的随机变量 x **用概率密度函数** (probability density function, pdf) $f(x)$ 或 $p(x)$ 来进行量化, 这些函数必须满足下表的条件:

特点	随机变量 x	
	离散	连续
取值范围	$x = a, a+1, \cdots, b$	$a \leqslant x \leqslant b$
pdf 的条件	$p(x) \geqslant 0, \sum\limits_{x=a}^{b} p(x) = 1$	$f(x) \geqslant 0, \int_a^b f(x)\mathrm{d}x = 1$

概率密度函数 $p(x)$ 或 $f(x)$ 必须是非负的 (否则, 某些事件概率的可能为负). 并且整个样本空间的概率必须等于 1.

累积分布函数 (cumulative distribution function, CDF) 是一个重要的概率度量, 定义为

$$\mathrm{P}\{x \leqslant X\} = \begin{cases} \mathrm{P}(X) = \sum\limits_{x=a}^{X} p(x), & x \text{ 离散} \\ F(X) = \int_a^X f(x)\mathrm{d}x, & x \text{ 连续} \end{cases}$$

例 14.2-1

考虑掷均匀骰子的情形. 随机变量 $x = \{1,2,3,4,5,6\}$ 表示骰子朝上的点数, 相应的 pdf 和 CDF 为

$$p(x) = \frac{1}{6}, x = 1, 2, \cdots, 6$$

$$\mathrm{P}(X) = \frac{X}{6}, X = 1, 2, \cdots, 6$$

图 14.1 画出了这两个函数. 概率密度函数 $p(x)$ 是一个**均匀离散函数** (uniform discrete function), 因为这个随机变量的所有取值的概率相等.

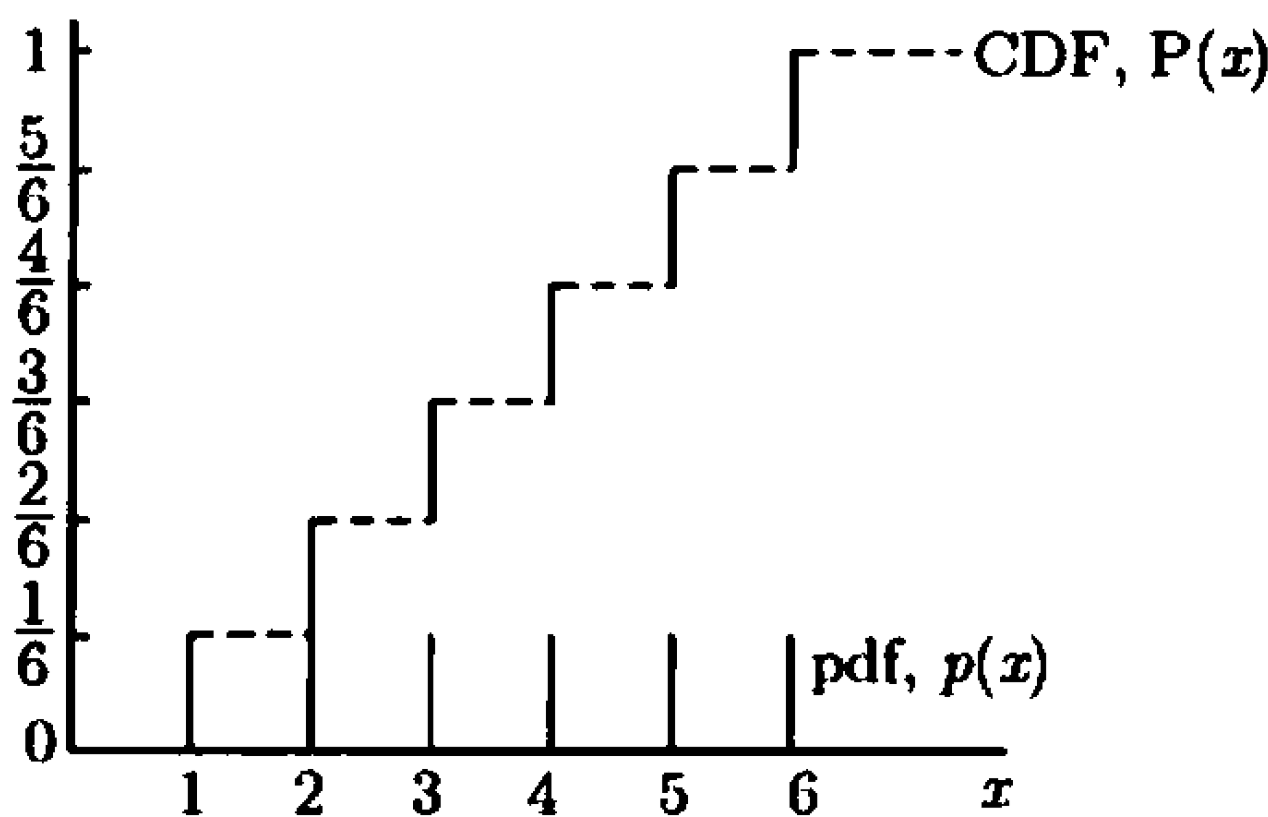


图 14.1 掷一枚骰子的 CDF 和 pdf

用下面的实验来说明均匀 $p(x)$ 的连续情形. 一根长度为 l 的针在一个直径也等于 l 的圆圈中心旋转, 在圆周上标注出任意一个参考点后, 我们按顺时针方向转动这根针, 并量出从指针停止点到标记点的圆周距离 x. 这样, 随机变量 x 在区间

$0 \leqslant x \leqslant \pi l$ 内是连续的. 没有理由相信, 这根针会有更多的时候停止在圆周的某个特定的区域. 因此, x 在指定区间的所有值都是等可能出现的, 因此 x 的分布一定是均匀的.

x 的概率密度函数 $f(x)$ 定义为

$$f(x) = \frac{1}{\pi l}, \quad 0 \leqslant x \leqslant \pi l$$

相应的累积分布函数 $F(X)$ 的计算公式为

$$F(X) = \mathrm{P}(x \leqslant X) = \int_0^X f(x)\mathrm{d}x = \int_0^X \frac{1}{\pi l}\mathrm{d}x = \frac{X}{\pi l}, \quad 0 \leqslant X \leqslant \pi l$$

图 14.2 画出了这两个函数的图形.

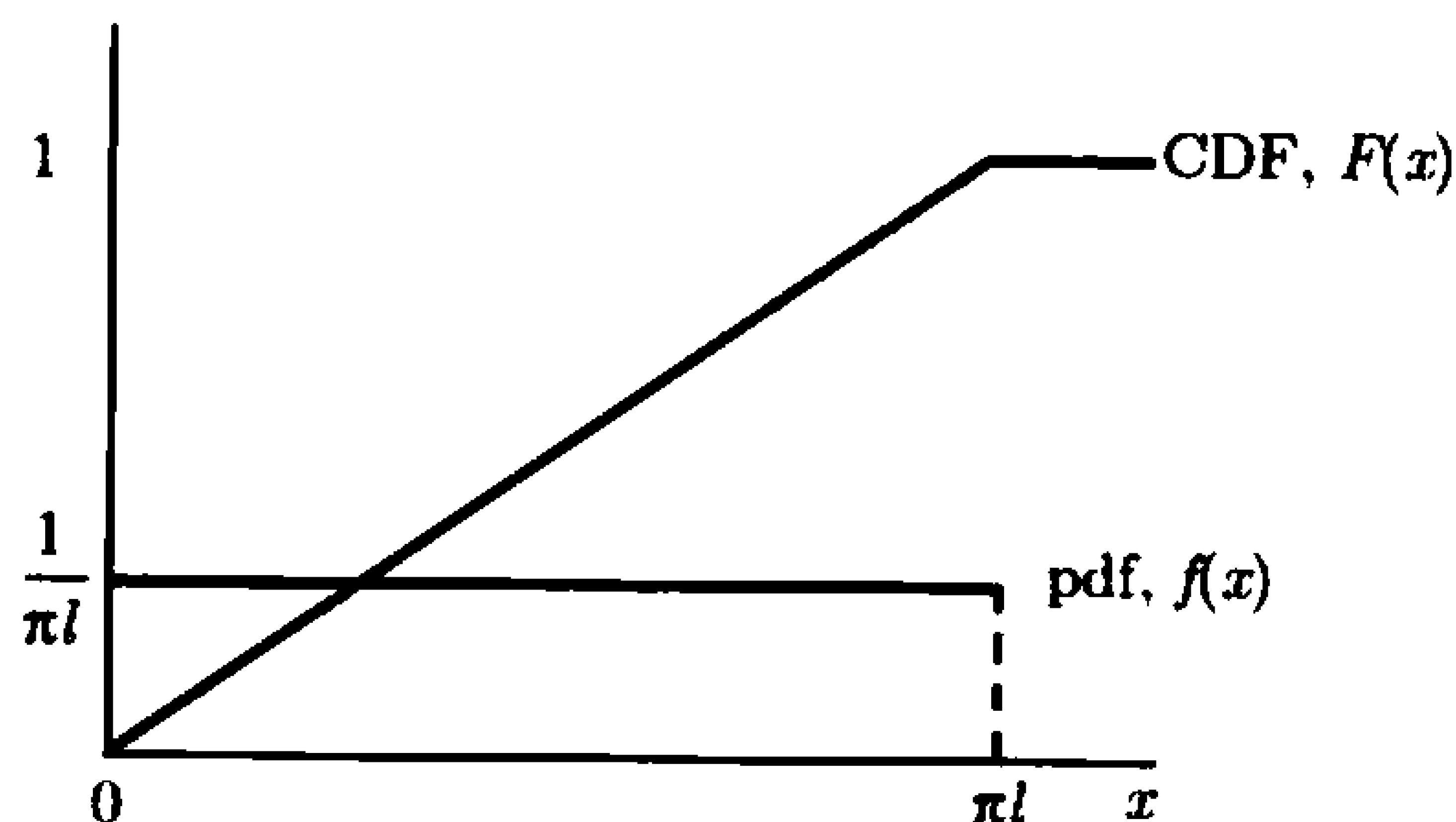


图 14.2 旋转针的 CDF 和 pdf

习题 14.2A

1. 某种物品所需要的件数 x 是 1~5 的离散数, 概率 $p(x)$ 与所需要的件数直接成比例, 比例常数为 K.
   (a) 给出 x 的 pdf 和 CDF, 画出这两个函数的图形.
   (b) 求出 x 是偶数值的概率.
2. 考虑下面的函数:

$$f(x) = \frac{k}{x^2}, \quad 10 \leqslant x \leqslant 20$$

   *(a) 求常数 k, 使得 $f(x)$ 为一个 pdf.
   (b) 给出 CDF 并分别求出 (i) x 大于 12, (ii) x 介于 13~15 的概率.

*3. 无铅汽油的日需求量服从 750~1 250 加仑之间的均匀分布, 容量为 1 100 加仑的汽油罐每天半夜添满. 该油罐刚好在补充前用完的概率是多少?

14.3 随机变量的期望

如果 $h(x)$ 为随机变量 x 的一个实函数, 我们定义 $h(x)$ 的**期望值** (expected value) $\mathrm{E}\{h(x)\}$ 为对 x 的 pdf 的 (长期) 加权平均. 用数学语言表示即是, 设 $p(x)$

和 $f(x)$ 分别为 x 的离散和连续的概率分布函数, 则 $\mathrm{E}\{h(x)\}$ 等于

$$\mathrm{E}\{h(x)\} = \begin{cases} \sum\limits_{x=a}^{b} h(x)p(x), & x \text{ 离散} \\ \int_a^b h(x)f(x)\mathrm{d}x, & x \text{ 连续} \end{cases}$$

例 14.3-1

(像许多人一样) 我每个月的第一周要支付所有的账单, 并回复几封信. 为此, 我每月需要买 20 枚邮票. 要用的邮票数量是 10~24 枚的等概率的随机数, 那么没用完剩下邮票的平均数是多少呢?

所用邮票枚数的 pdf 是

$$p(x) = \frac{1}{15}, \quad x = 10, 11, \cdots, 24.$$

剩下的邮票数就是

$$h(x) = \begin{cases} 20 - x, & x = 10, 11, \cdots, 19 \\ 0, & \text{其他情形} \end{cases}$$

因此

$$\begin{aligned} \mathrm{E}\{h(x)\} &= \frac{1}{15}[(20-10) + (20-11) + (20-12) + \cdots + (20-19)] + \frac{5}{15}(0) \\ &= 3\frac{2}{3} \end{aligned}$$

我们需要乘积 $\frac{5}{15}(0)$ 来计算出完整的 $h(x)$ 的期望值. 特别地, 剩下 0 枚多余邮票的概率就等于需要用 20 枚或以上邮票的概率, 即

$$\mathrm{P}\{x \geqslant 20\} = p(20) + p(21) + p(22) + p(23) + p(24) = 5\left(\frac{1}{15}\right) = \frac{5}{15}$$

习题 14.3A

1. 在例 14.3-1 中, 计算出为满足你最大可能的需求, 另外还需要的邮票平均枚数.
2. 例 14.3-1 和第 1 题的结果表明, 多余和不够的邮票的平均数都是正的. 这些结果有什么矛盾吗? 请解释.

*3. 某报摊每天早上收到 50 份*Al Ahram*报, 每天卖出的份数 x 根据下面的概率分布随机变化:

$$p(x) = \begin{cases} \frac{1}{45}, & x = 35, 36, \cdots, 49 \\ \frac{1}{30}, & x = 50, 51, \cdots, 59 \\ \frac{1}{33}, & x = 60, 61, \cdots, 70 \end{cases}$$

(a) 求报摊卖出所有报纸的概率.

(b) 求每天没卖出去的报纸的平均数.

(c) 假如报摊花 50 美分买进一份报纸, 以 \$1.00 一份卖出, 求报摊每天的期望净收入.

14.3.1 随机变量的平均值和方差 (标准差)

x **的均值** (mean) $\mathrm{E}\{x\}$ 是对该随机变量中心趋势 (或加权和) 的一种数值度量, **而方差** (variance) $\mathrm{var}\{x\}$ 衡量 x 围绕平均值 $\mathrm{E}\{x\}$ 的离差或偏差, 它的平方根则称为 x 的**标准差** (standard deviation) $\mathrm{stdDev}\{x\}$. 标准差越大, 就意味着关于该随机变量的不确定性程度越高. 特别地, 当一个变量的值等于某一常量时, 它的标准差为零.

计算平均值和方差的公式可以从 $\mathrm{E}\{h(x)\}$ 的一般定义得出: 对 $\mathrm{E}\{x\}$, 用 $h(x) = x$; 对于 $\mathrm{var}\{x\}$, 用 $h(x) = (x - \mathrm{E}\{x\})^2$. 因此

$$\mathrm{E}\{x\} = \begin{cases} \displaystyle\sum_{x=a}^{b} xp(x), & x \text{ 离散} \\ \displaystyle\int_a^b xf(x)\mathrm{d}x, & x \text{ 连续} \end{cases}$$

$$\mathrm{var}\{x\} = \begin{cases} \displaystyle\sum_{x=a}^{b} (x - \mathrm{E}\{x\})^2 p(x), & x \text{ 离散} \\ \displaystyle\int_a^b (x - \mathrm{E}\{x\})^2 f(x)\mathrm{d}x, & x \text{ 连续} \end{cases}$$

$$\mathrm{stdDev}\{\mathrm{x}\} = \sqrt{\mathrm{var}\{x\}}$$

通过对离散情况的考察, 可以更容易地理解这些公式的原理. 这里, $\mathrm{E}\{x\}$ 是 x 离散值的*加权和*, $\mathrm{var}\{x\}$ 也是围绕 $\mathrm{E}\{x\}$ 偏差平方的*加权和*. 可以类似地解释连续的情况, 只是要用积分来代替求和.

例 14.3-2

计算例 14.2-1 中两个实验的平均值和方差.

实验 1(掷骰子) pdf 为 $p(x) = \frac{1}{6}$, $x = 1, 2, \cdots, 6$. 因此

$$\begin{aligned} \mathrm{E}\{x\} &= 1\left(\tfrac{1}{6}\right) + 2\left(\tfrac{1}{6}\right) + 3\left(\tfrac{1}{6}\right) + 4\left(\tfrac{1}{6}\right) + 5\left(\tfrac{1}{6}\right) + 6\left(\tfrac{1}{6}\right) = 3.5 \\ \mathrm{var}\{x\} &= \left(\tfrac{1}{6}\right)\left\{(1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 \right. \\ &\quad \left. + (5-3.5)^2 + (6-3.5)^2\right\} = 2.917 \\ \mathrm{stdDev}\{x\} &= \sqrt{2.917} = 1.708 \end{aligned}$$

实验 2(旋转针) 假设针的长度为 1 英寸, 则

$$f(x) = \frac{1}{3.14}, \quad 0 \leqslant x \leqslant 3.14$$

平均值和方差可计算为

$$\mathrm{E}\{x\} = \int_0^{3.14} x\left(\frac{1}{3.14}\right)\mathrm{d}x = 1.57 \text{ 英寸}$$

$$\text{var}\{x\}=\int_0^{3.14}(x-1.57)^2\left(\frac{1}{3.14}\right)\mathrm{d}x=0.822\ 英寸^2$$

$$\text{stdDev}\{x\}=\sqrt{0.822}=0.906\ 英寸$$

Excel 程序

电子表格程序 exelStatTables.xls 用来计算 16 个常用 pdf(包括例 14.3-2 的离散和连续均匀分布) 的平均值、标准差、概率和百分位数. 程序本身是非常浅显易懂的.

习题 14.3B

*1. 计算习题 14.2A 第 1 题中定义的随机变量的平均值和方差.

2. 计算习题 14.2A 第 2 题的随机变量的平均值和方差.

3. 证明一个均匀随机变量 x, $a \leqslant x \leqslant b$ 的平均值和方差为

$$\text{E}\{x\}=\frac{b+a}{2}$$

$$\text{var}\{x\}=\frac{(b-a)^2}{12}$$

4. 对给定的 pdf $f(x)$, $a \leqslant x \leqslant b$, 证明

$$\text{var}\{x\}=\text{E}\{x^2\}-(\text{E}\{x\})^2$$

5. 对给定的 pdf $f(x)$, $a \leqslant x \leqslant b$, 以及 $y=cx+d$, 其中 c 和 d 为常数, 证明

$$\text{E}\{y\}=c\,\text{E}\{x\}+d$$

$$\text{var}\{y\}=c^2\,\text{var}\{x\}$$

14.3.2 联合随机变量的平均值和方差

考虑两个连续的随机变量 x_1 和 x_2, 其中 $a_1 \leqslant x_1 \leqslant b_1$, $a_2 \leqslant x_2 \leqslant b_2$. 定义 $f(x_1, x_2)$ 是 x_1 和 x_2 的**联合概率密度函数** (joint probability density function), 且 $f_1(x_1)$, $f_2(x_2)$ 分别为 x_1 和 x_2 的**边际概率密度函数** (marginal probability density function), 则

$$f(x_1,x_2)\geqslant 0, a_1\leqslant x_1\leqslant b_1, a_2\leqslant x_2\leqslant b_2$$

$$\int_{a_1}^{b_1}\mathrm{d}x_1\int_{a_2}^{b_2}\mathrm{d}x_2 f(x_1,x_2)=1$$

$$f_1(x_1)=\int_{a_2}^{b_2}f(x_1,x_2)\mathrm{d}x_2$$

$$f_2(x_2)=\int_{a_1}^{b_1}f(x_1,x_2)\mathrm{d}x_1$$

$$f(x_1, x_2) = f_1(x_1)f_2(x_2), \text{ 如果 } x_1, x_2 \text{ 是独立的}$$

这些公式同样适用于离散的 pdf, 只要把积分换成求和.

对于 $y = c_1x_1 + c_2x_2$ 的特殊情况, 其中随机变量 x_1 和 x_2 服从 pdf $f(x_1, x_2)$ 的联合分布, 可以证明

$$\mathrm{E}\{c_1x_1 + c_2x_2\} = c_1\mathrm{E}\{x_1\} + c_2\mathrm{E}\{x_2\}$$

$$\mathrm{var}\{c_1x_1 + c_2x_2\} = c_1^2\mathrm{var}\{x_1\} + c_2^2\mathrm{var}\{x_2\} + 2c_1c_2\mathrm{cov}\{x_1, x_2\}$$

其中

$$\begin{aligned}\mathrm{cov}\{x_1, x_2\} &= \mathrm{E}\{(x_1 - \mathrm{E}\{x_1\})(x_2 - \mathrm{E}\{x_2\})\} \\ &= \mathrm{E}(x_1x_2 - x_1\mathrm{E}\{x_2\} - x_2\mathrm{E}\{x_1\} + \mathrm{E}\{x_1\}\mathrm{E}\{x_2\}) \\ &= \mathrm{E}\{x_1x_2\} - \mathrm{E}\{x_1\}\mathrm{E}\{x_2\}\end{aligned}$$

若 x_1, x_2 独立, 则 $\mathrm{E}\{x_1, x_2\} = \mathrm{E}\{x_1\}\mathrm{E}\{x_2\}$ 且 $\mathrm{cov}\{x_1, x_2\} = 0$. 反之则不然, 因为两个相关的随机变量的协方差也可能为零.

例 14.3-3

一批产品包括 4 件次品 (D) 和 6 件合格品 (G). 随机选择一件进行检验, 在不放回情况下再检验第 2 件产品. 令随机变量 x_1 和 x_2 分别表示第 1 件和第 2 件产品的抽样检验结果.

(1) 求 x_1 和 x_2 的联合概率密度函数和边际概率密度函数.

(2) 假定对每件选中的合格品你得到 \$5, 但如果是次品你必须付 \$6, 求在两件产品选定后你的收益的平均值和方差.

令 $p(x_1, x_2)$ 为 x_1 和 x_2 的联合概率密度函数, 定义 $p_1(x_1)$, $p_2(x_2)$ 为分别的边际概率密度函数. 首先求 $p_1(x_1)$ 如下:

$$p_1(\mathrm{G}) = \tfrac{6}{10} = 0.6, \quad p_1(\mathrm{D}) = \tfrac{4}{10} = 0.4$$

接下来, 我们知道第二个结果 x_2 依赖于 x_1, 因此为了确定 $p_2(x_2)$, 首先求出联合概率密度函数 $p(x_1, x_2)$, 由此可以确定边际分布 $p_2(x_2)$.

$$\begin{aligned}\mathrm{P}\{x_2 = \mathrm{G}|x_1 = \mathrm{G}\} &= \frac{5}{9} \\ \mathrm{P}\{x_2 = \mathrm{G}|x_1 = \mathrm{B}\} &= \frac{6}{9} \\ \mathrm{P}\{x_2 = \mathrm{B}|x_1 = \mathrm{G}\} &= \frac{4}{9} \\ \mathrm{P}\{x_2 = \mathrm{B}|x_1 = \mathrm{B}\} &= \frac{3}{9}\end{aligned}$$

为了确定 $p(x_1,x_2)$, 利用公式 $P\{AB\}=P\{A|B\}P\{B\}$ (见 14.1.2 节).

$$p\{x_2=G,x_1=G\}=\frac{5}{9}\times\frac{6}{10}=\frac{5}{15}$$
$$p\{x_2=G,x_1=B\}=\frac{6}{9}\times\frac{4}{10}=\frac{4}{15}$$
$$p\{x_2=B,x_1=G\}=\frac{4}{9}\times\frac{6}{10}=\frac{4}{15}$$
$$p\{x_2=B,x_1=B\}=\frac{3}{9}\times\frac{4}{10}=\frac{2}{15}$$

可以用下列方法得到边际分布 $p_1(x_1)$ 和 $p_2(x_2)$: 把联合分布函数 $p(x_1,x_2)$ 表示成表的形式, 然后按下表所示分别将行列相加.

	$x_2=G$	$x_2=B$	$p_1(x_1)$
$x_1=G$	$\frac{5}{15}$	$\frac{4}{15}$	$\frac{9}{15}=0.6$
$x_1=B$	$\frac{4}{15}$	$\frac{2}{15}$	$\frac{6}{15}=0.4$
$p_2(x_2)$	$\frac{9}{15}=0.6$	$\frac{6}{15}=0.4$	

有意思的是, $p_1(x_1)=p_2(x_2)$, 这跟我们的直觉正好相反.

依题意, 由 G 得到 \$5, 由 B 得到 −\$6, 那么从联合分布可以确定期望收益.

$$\text{期望收益}=(5+5)\frac{5}{15}+(5-6)\frac{4}{15}+(-6+5)\frac{4}{15}+(-6-6)\frac{2}{15}=\$1.20$$

根据两次选择的期望收益等于每次单独选择的期望收益之和 (虽然这两个变量不是独立的), 可以得出同样的结果, 这意味着

$$\begin{aligned}\text{期望收益}&=\text{选择产品 1 的期望收益}+\text{选择产品 2 的期望收益}\\&=(5\times0.6-6\times0.4)+(5\times0.6-6\times0.4)=\$1.20\end{aligned}$$

为了计算总收益的方差, 注意到

$$\text{var}\{\text{收益}\}=\text{var}\{\text{收益 1}\}+\text{var}\{\text{收益 2}\}+2\text{cov}\{\text{收益 1, 收益 2}\}$$

由于 $p_1(x_1)=p_2(x_2)$, 得到 var{收益 1} = var{收益 2}. 为计算方差, 用公式

$$\text{var}\{x\}=E\{x^2\}-(E\{x\})^2$$

(见习题 14.3B 第 4 题.) 因此

$$\text{var}\{\text{收益 1}\}=[5^2\times0.6+(-6)^2\times0.4]-0.6^2=29.04$$

接下来, 为了计算协方差, 用公式

$$\text{cov}\{x_1,x_2\}=E\{x_1x_2\}-E\{x_1\}E\{x_2\}$$

$E\{x_1x_2\}$ 一项可以从 x_1 和 x_2 的联合概率密度函数求出, 因此有

$$\text{协方差} = (5\times5)(\tfrac{5}{15}) + [5\times(-6)](\tfrac{4}{15}) + [(-6)\times5](\tfrac{4}{15}) + [(-6)\times(-6)](\tfrac{2}{15}) - 0.6\times0.6 = -3.23$$

因此

$$\text{方差} = 29.04 + 29.04 + 2(-3.23) = 51.62$$

习题 14.3C

1. x_1 和 x_2 的联合 pdf $p(x_1, x_2)$ 为

	$x_2=3$	$x_2=5$	$x_2=7$
$x_1=1$	0.2	0	0.2
$x_1=2$	0	0.2	0
$x_1=3$	0.2	0	0.2

*(a) 求边际 pdf $p_1(x_1)$ 和 $p_2(x_2)$.

*(b) x_1 和 x_2 独立吗?

(c) 计算 $E\{x_1+x_2\}$.

(d) 计算 $\text{cov}\{x_1, x_2\}$.

(e) 计算 $\text{var}\{5x_1-6x_2\}$.

14.4 4 种常用概率分布

14.2 节和 14.3 节讨论了 (离散的和连续的) 均匀分布, 本节介绍另外 4 个运筹学中常见的概率密度函数: 离散的二项分布和泊松分布, 以及连续的指数分布和正态分布.

14.4.1 二项分布

假定一家制造公司生产某种产品, 每批量生产 n 件, 每个批次中不合格产品的比例 p 可以从历史数据中估算出来. 我们感兴趣的是一个批次中不合格产品数量的 pdf.

一批 n 件产品中有 x 个不合格品的不同组合有 $C_n^x = \frac{n!}{x!(n-x)!}$ 种, 并且得到每种组合的概率是 $p^x(1-p)^{n-x}$. 因此得出 (根据概率的加法律) 一批 n 件产品中有 k 件不合格品的概率为

$$P\{x=k\} = C_n^k p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \cdots, n$$

这就是带有参数 n 和 p 的二项分布, 其平均值和方差为

$$E\{x\} = np \qquad \text{var}\{x\} = np(1-p)$$

例 14.4-1

John Doe 的工作需要每天在两个小镇之间开车往返 10 次, 完成了这所有的 10 趟活, Doe 先生就可以歇息了, 这让他老是想开快车. 根据他过去的经历, 每趟往返有 40% 的机会收到超速罚款单.

(1) 当天结束时未收到超速罚单的概率是多少?

(2) 若每次超速罚款 \$80, 每天平均罚款额有多少?

任何一次往返收到一张罚单的概率是 $p = 0.4$, 因此一天内没有罚款的概率就是

$$\mathrm{P}\{x = 0\} = C_{10}^{0}(0.4)^{0}(0.6)^{10} = 0.006$$

这就是说完成一天的工作而没受到罚款的机会不到 1%. 事实上, 每天平均罚款额可计算如下:

$$\text{平均罚款额} = \$80\mathrm{E}\{x\} = \$80(np) = 80 \times 10 \times 0.4 = \$320$$

评注 $\mathrm{P}\{x = 0\}$ 可以用 excelStatTables.xls 程序计算. 在 F7 中输入 10, 在 G7 输入 0.4, 在 J7 中输入 0. 答案 $\mathrm{P}\{x = 0\} = 0.006\,047$ 就会在 M7 中给出.

习题 14.4A

***1.** 一只均匀骰子掷 10 次, 掷出的骰子不出现偶数的概率是多少?

2. 设有 5 枚均匀硬币独立地抛掷, 恰好有 1 枚硬币与其他 4 枚不一样的概率是多少?

***3.** 一个算卦先生称, 他能通过观察笔迹预测出人们一生积聚财富的能力. 为了检验他是否灵验, 10 个百万富翁和 10 位大学教授提供了他们笔迹的样本, 然后把这些样本凑成对, 每对样本由一个富翁一个教授组成, 并交给算命先生看. 假如这位算命先生作出至少 8 次正确的预测, 我们就说他是灵验的. 那么由"侥幸说对"而灵验的概率是多少?

4. 在赌场的一个赌博游戏中, 要求你在服务员同时投掷 3 个均匀骰子之前选择 1~6 的一个数字. 只要有骰子的点数跟你选的数一样, 赌场根据你选对的骰子个数付给你钱. 如果都选错了, 你只付给赌场 \$1. 你从这项游戏中获得的长期期望收益是多少?

5. 假定你正在玩下面的游戏: 掷 2 只均匀骰子, 假如点数不同, 你付出 10 美分; 若点数相同, 你得到 50 美分. 这项游戏的期望收益如何?

6. 证明二项分布的平均值和方差公式.

14.4.2 泊松分布

顾客按照某种"完全随机"的方式到达一家银行或杂货店, 也就是说, 我们无法预测什么时候有人到达. 描述某个具体期间这种到达数 的 pdf 就是泊松分布.

令 x 为某个具体时间单元 (如一分钟或一小时) 期间发生的事件 (如到达) 的次数. 假定 λ 是一个已知的常数, 泊松分布的 pdf 定义为

$$\mathrm{P}\{x = k\} = \frac{\lambda^{k}\mathrm{e}^{-\lambda}}{k!}, k = 0, 1, 2, \cdots$$

泊松分布的平均值和方差为

$$E\{x\} = \lambda \qquad \text{var}\{x\} = \lambda$$

平均值公式显示, λ 代表事件发生的频率.

泊松分布常用来描述随机排队服务系统 (见第 12 章).

例 14.4-2

一家小型机器修理车间的维修任务完全随机地到达, 每天 10 次.

(1) 车间每天接到的平均维修任务数是多少?

(2) 假设车间每天营业 8 小时, 那么任何 1 小时期间没有维修任务的概率是多少?

每天接到的任务平均数 $\lambda = 10$. 为了计算每小时无任务到达的概率, 我们需要计算每小时的到达率, 即每小时任务率 $\lambda_h = \frac{10}{8} = 1.25$. 因此

$$P\{\text{任一小时无任务到达}\} = \frac{(\lambda_h)^0 e^{-\lambda_h}}{0!} = \frac{1.25^0 e^{-1.25}}{0!} = 0.286\,5$$

评注 上述概率可以用 excelStatTables.xls 计算. 在 F16 中输入 1.25, 在 J16 输入 0. 则在 M16 中显示答案 0.286 505.

习题 14.4B

*1. 已知顾客按照泊松分布来到某服务机构, 到达率为每分钟 4 人. 求在任何给定的 30 秒区间内至少有一位顾客到达的概率.

2. 参数为 λ 的泊松分布接近于参数为 (n, p) 当 $n \to \infty, p \to 0, np \to \lambda$ 的二项分布. 针对某个制造批量包含 1% 次品的情形证明该结论. 假如从批量中抽出 10 件样品, 计算在样品中至多有一件次品的概率, 先用 (精确) 二项分布然后用 (近似) 泊松分布计算. 证明当 p 值增加到 0.5 时, 该近似是不可接受的.

*3. 顾客以每小时 20 人的频率随机来到某检查站.
(a) 求该检查站空闲的概率. (b) 求至少有 2 个人排队等待服务的概率.

4. 证明泊松分布的平均值和方差公式.

14.4.3 负指数分布

如某个服务设施在指定期间的到达数服从泊松分布 (见 14.4.2 节), 则自动地, 两个连续到达的时间间隔的分布必服从负指数分布 (简称指数分布). 特别地, 如果 λ 是泊松事件的发生率, 则两次连续到达的间隔时间分布 x 为

$$f(x) = \lambda e^{-\lambda x}, x > 0$$

图 14.3 给出了 $f(x)$ 的图像.

指数分布的平均值和方差为

$$E\{x\} = \frac{1}{\lambda} \qquad \text{var}\{x\} = \frac{1}{\lambda}$$

平均值 E$\{x\}$ 与 λ 的定义一致. 若 λ 为事件的发生率, 则 $\frac{1}{\lambda}$ 是两次连续事件的平均间隔时间.

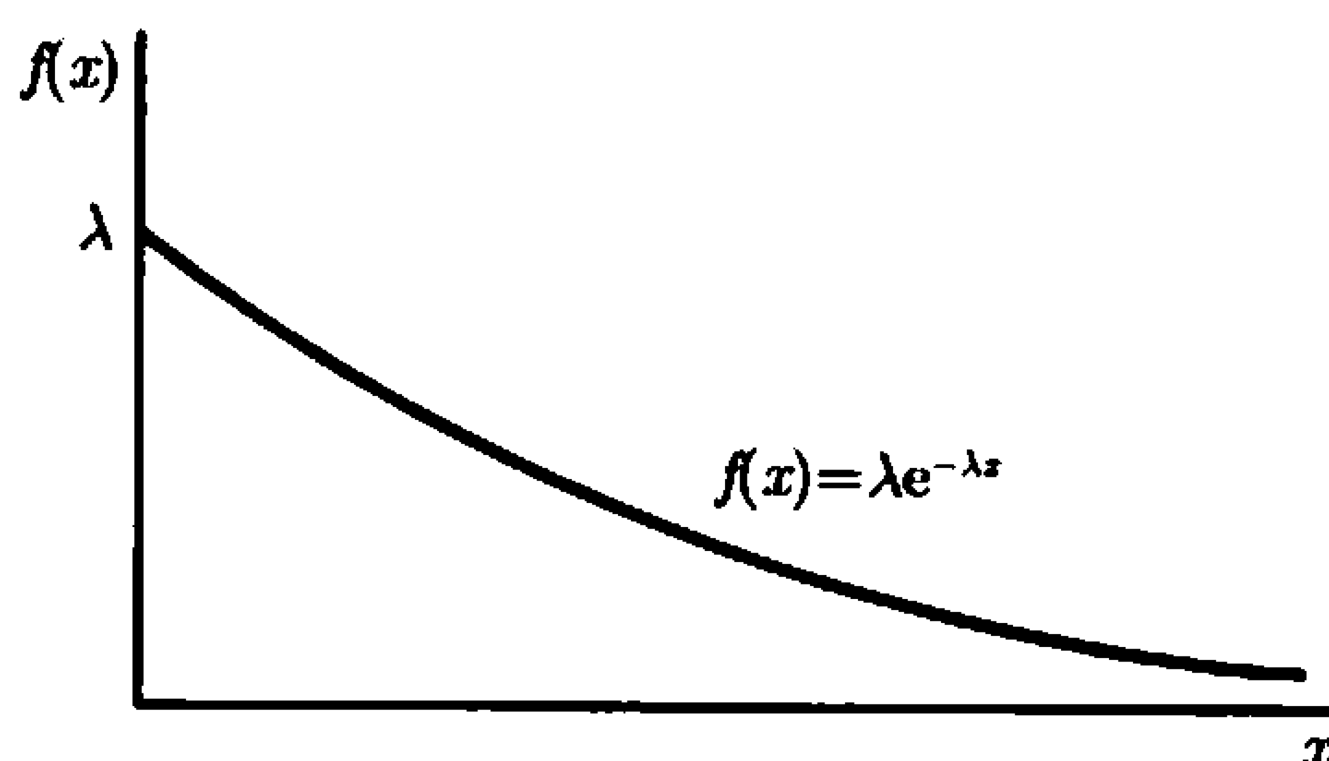


图 14.3 指数分布的概率密度函数

例 14.4-3

汽车平均每两分钟一辆随机来到某加油站. 求车辆到达间隔时间不超过 1 分钟的概率.

这个例子所求的概率形式为 P$\{x \leqslant A\}$, 其中 $A = 1$ 分钟. 求这个概率和计算 x 的 CDF 是一样的, 即

$$\mathrm{P}\{x \leqslant A\} = \int_0^A \lambda \mathrm{e}^{-\lambda x}\mathrm{d}x = -\mathrm{e}^{-\lambda x}\big|_0^A = 1 - \mathrm{e}^{-\lambda A}$$

该例中到达率可计算为

$$\lambda = \frac{1}{2}\text{辆/分钟}$$

因此,

$$\mathrm{P}\{x \leqslant 1\} = 1 - \mathrm{e}^{-(\frac{1}{2})(1)} = 0.3934$$

评注 可以用 excelStatTables.xls 计算上述概率. 在 F9 输入 0.5, 在 J9 输入 1, O9 会给出答案 0.393 468.

习题 14.4C

*1. 沃尔玛店的顾客来自城里和附近的郊区. 城里顾客的到达率为每分钟 5 人, 郊区顾客每分钟来 7 人, 这些人随机到达. 求所有顾客到达间隔时间不到 5 秒钟的概率.

2. 证明指数分布的平均值和方差公式.

14.4.4 正态分布

正态分布描述了我们日常生活中发生的许多现象, 包括考试分数、体重、身高, 等等. 正态分布的 pdf 定义为

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

均值和方差为

$$E\{x\} = \mu$$
$$\mathrm{var}\{x\} = \sigma^2$$

通常用符号 $N(\mu, \sigma)$ 来表示平均值为 μ、标准差为 σ 的正态分布.

图 14.4 给出了正态 pdf $f(x)$ 的图像. 这个函数关于平均值 μ 总是对称的.

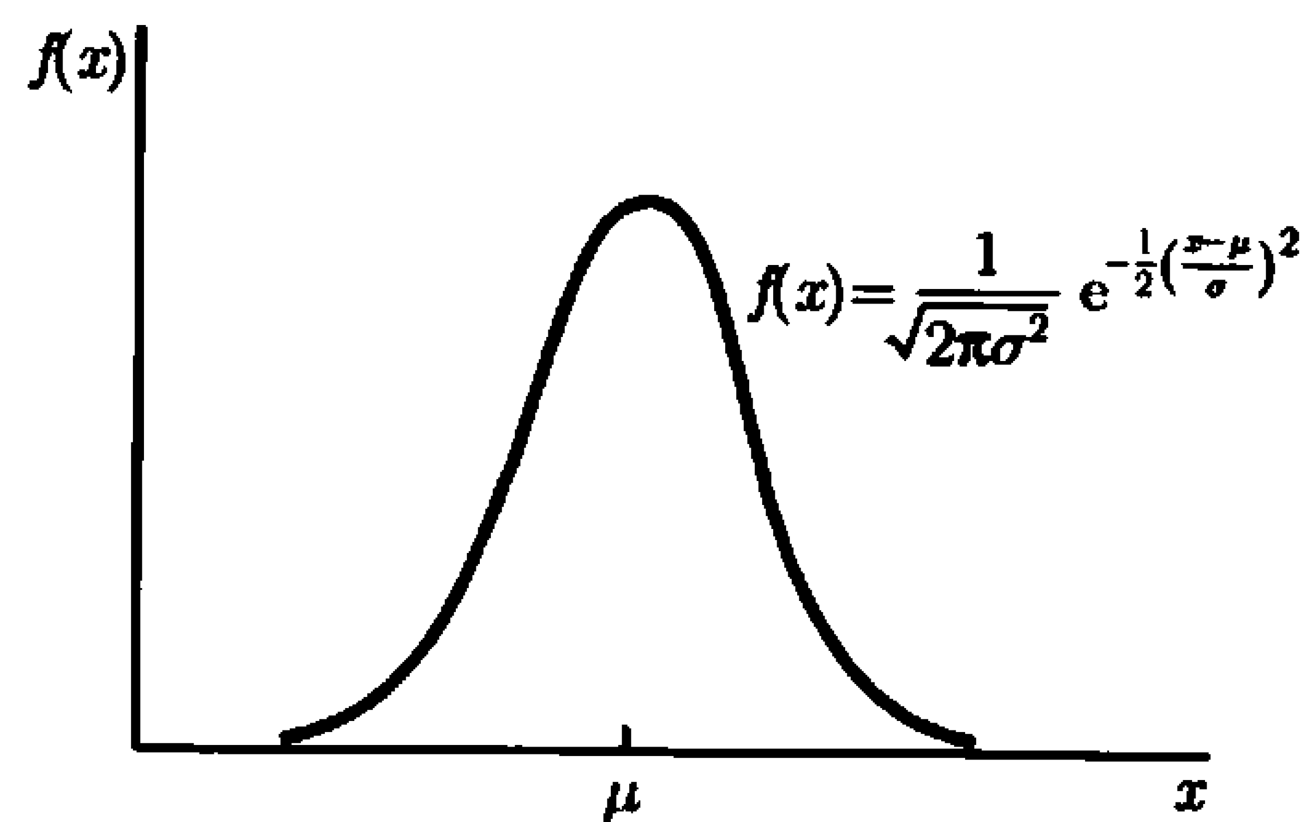


图 14.4 正态随机变量的概率密度函数

正态随机变量的一条重要性质是, 它近似于来自任何分布样本平均值的分布. 这一重要的结果出自以下定理:

中心极限定理 令 $x_1, x_2, \cdots, x_n$ 为独立同分布的随机变量, 每个变量的平均值为 μ, 标准差为 σ, 定义

$$s_n = x_1 + x_2 + \cdots + x_n$$

随着 n 变大 $(n \to \infty)$, 不论 $x_1, x_2, \cdots, x_n$ 原来分布为何, s_n 的分布总是趋向于均值为 $n\mu$、方差为 $n\sigma^2$ 的正态分布.

中心极限定理特别告诉我们, 来自任何分布的 n 个样本的均值的分布都是渐进正态的, 其均值是 μ, 方差为 $\frac{\sigma^2}{n}$. 这一结果在统计质量控制中有着重要的应用.

正态随机变量的 CDF 不能以闭形式求出. 因此, 我们编制了正态表 (附录 B 表 1, 或者 excelStatTables.xls), 这些表适用于平均值是 0、标准差为 1 的**标准正态**分布, 即 $N(0,1)$. 任何正态随机变量 x(均值为 μ, 标准差为 σ) 都可以利用以下变换转换成标准正态:

$$z = \frac{x - \mu}{\sigma}.$$

任何正态分布下, 99% 的面积都含在区间 $\mu - 3\sigma \leqslant x \leqslant \mu + 3\sigma$ 之内, 称为6σ **限制**.

例 14.4-4

某个圆筒的内径尺寸要求为 1 ± 0.03 cm. 机械加工结果服从均值为 1 cm, 标准差 0.1 cm 的正态分布. 求出产品符合工艺要求的百分比.

令 x 表示加工结果, 一只圆筒符合加工要求的概率为

$$P\{1 - 0.03 \leqslant x \leqslant 1 + 0.03\} = P\{0.97 \leqslant x \leqslant 1.03\}$$

已知 $\mu = 1, \sigma = 0.1$, 等价的标准正态概率陈述就是

$$
\begin{aligned}
P\{0.97 \leqslant x \leqslant 1.03\} &= P\left\{\frac{0.97-1}{0.1} \leqslant z \leqslant \frac{1.03-1}{0.1}\right\} = P\{-0.3 \leqslant z \leqslant 0.3\} \\
&= P\{z \leqslant 0.3\} - P\{z \leqslant -0.3\} = P\{z \leqslant 0.3\} - P\{z \geqslant 0.3\} \\
&= P\{z \leqslant 0.3\} - [1 - P\{z \leqslant 0.3\}] \\
&= 2P\{z \leqslant 0.3\} - 1 = 2 \times 0.617\,9 - 1 = 0.235\,8
\end{aligned}
$$

上述概率结果可以用图 14.5 中的阴影区域得到印证. 注意到由于 pdf 的对称性, $P\{z \leqslant -0.3\} = 1 - P\{z \leqslant 0.3\}$, 从标准正态表 (附录 B 中表 1) 可查出值 $0.617\,9 (= P\{z \leqslant 0.3\})$.

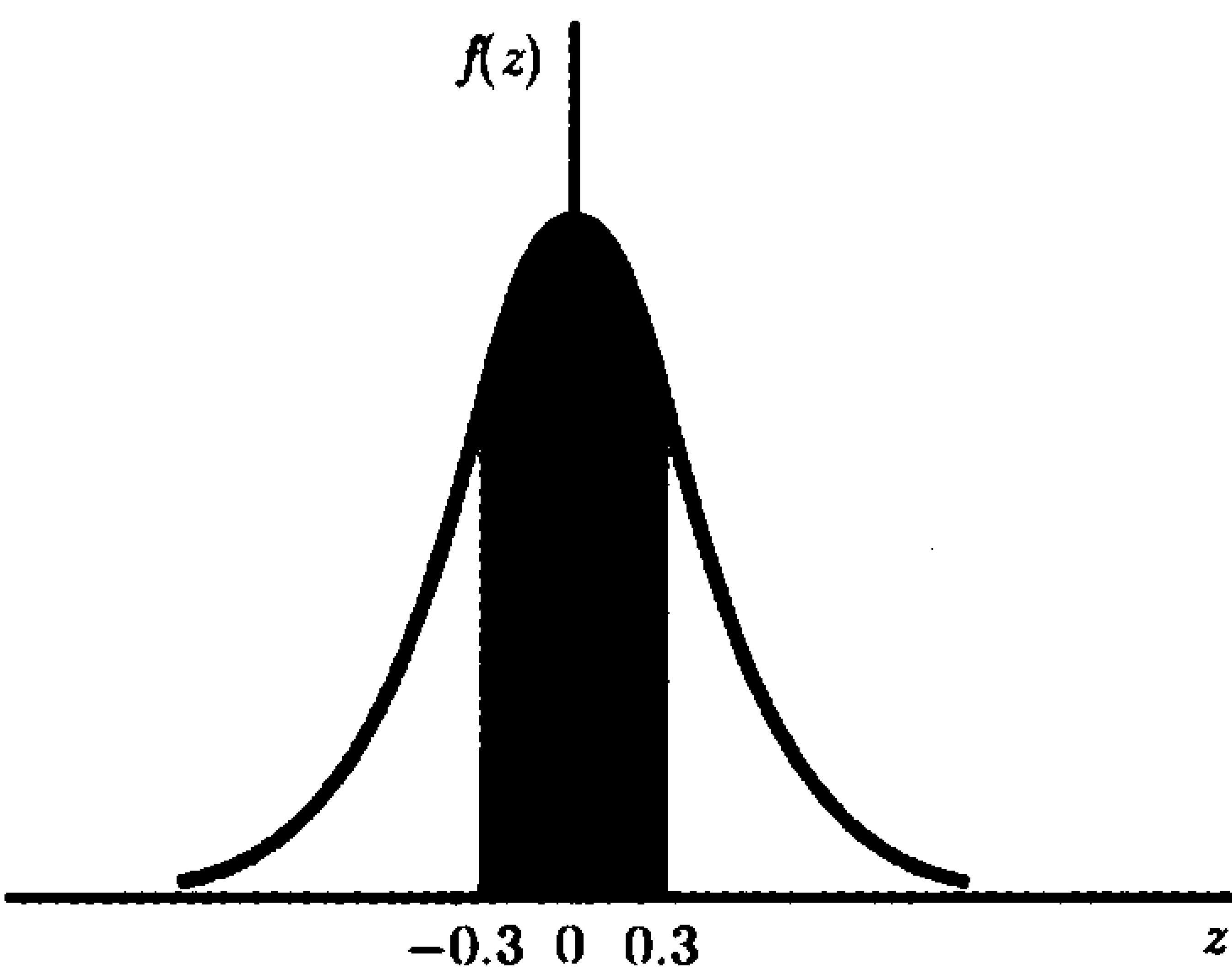


图 14.5 标准正态分布中计算 $P\{-0.3 \leqslant z \leqslant 0.3\}$

评注 $P\{a \leqslant x \leqslant b\}$ 可以用 excelStatTables.xls 直接计算. 在 F15 输入 1, 在 G15 输入 0.1, 在 J15 输入 0.97, 在 K15 输入 1.03. Q15 会给出答案 0.235 823.

习题 14.4D

1. *A* 大学工程学院要求 ACT 成绩最少 26 分. 某校区高中三年级的考试成绩呈正态分布, 平均分为 22 分, 标准差为 2 分.
   (a) 求高三学生可能考入工程学院的百分比.
   (b) 假如 *A* 大学不接受 ACT 成绩低于 17 分的学生, 求不符合 *A* 大学入学条件的学生比例.

*2. 某个游乐园里想要乘坐直升飞机的人平均体重为 180 磅, 标准差为 15 磅. 该直升飞机可搭乘 5 人, 但最大载重量只有 1 000 磅. 该直升机上乘坐了 5 个人但不能起飞的概率是多少 (提示: 运用中心极限定理).

3. 某个圆筒的内径尺寸服从正态分布, 均值为 1 cm, 标准差 0.1 cm. 一个实心铁棒要安放在每个圆筒内. 该铁棒也服从正态分布, 均值为 0.99 cm, 标准差 0.01 cm. 求出安装的铁棒–圆筒对不能匹配的百分比. (提示: 两个正态分布随机变量的差也是正态的.)

14.5 经验分布

前面的几节介绍了随机变量的概率密度函数和累积密度函数, 并给出了 5 个常用分布的例子 (均匀分布、二项分布、泊松分布、指数分布和正态分布). 那么在实际中如何确定这些分布呢?

确定 (实际上是估计出) 任何一个 pdf, 主要根据特定情况下收集的原始数据. 例如, 要估计出一个杂货店顾客到达的间隔时间的 pdf, 首先要记录出顾客到达的具体时刻, 所要的间隔数据就是连续到达时刻之间的差值.

本节说明如何从采样数据转换到一个 pdf.

第 1 步 把原始数据用某个合适的频率直方图表示出来, 并确定出相应的经验 pdf.

第 2 步 利用*拟合优度检验*, 检查所得到的经验 pdf 是否为某个已知理论 pdf 的样本.

频率直方图 频率直方图是通过把原始数据用数据的范围 (从最小值到最大值) 相除, 得到一些互不重叠的小区间来构造的. 给定小区间 i 的边界 $(I_{i-1}, I_i]$, 对应的频数确定为所有满足 $I_{i-1} < x \leqslant I_i$ 的原始数据 x 的个数.

例 14.5-1

下表数据表示某个服务机构对 60 位顾客样本的服务时间 (分钟):

0.7	0.4	3.4	4.8	2.0	1.0	5.5	6.2	1.2	4.4
1.5	2.4	3.4	6.4	3.7	4.8	2.5	5.5	0.3	8.7
2.7	0.4	2.2	2.4	0.5	1.7	9.3	8.0	4.7	5.9
0.7	1.6	5.2	0.6	0.9	3.9	3.3	0.2	0.2	4.9
9.6	1.9	9.1	1.3	10.6	3.0	0.3	2.9	2.9	4.8
8.7	2.4	7.2	1.5	7.9	11.7	6.3	3.8	6.9	5.3

这批数据的最小值和最大值分别是 0.2 和 11.7. 这就意味着, 所有数据可以被范围 $(0,12]$ 所覆盖. 把区间 $(0,12]$ 任意分成 12 个小区间, 每个小区间宽度为 1 分钟. 适当选择小区间的宽度对于反映出经验分布的形状至关重要, 虽然对于确定最优的小区间宽度没有硬性的规则, 但根据经验我们通常采用 10~20 个小区间. 实际上, 有必要试验不同的小区间宽度, 以得到可以接受的直方图.

下表汇总了给定原始数据的直方图信息. 相对频数列 f_i 可由观测频数列 o_i 的数据个数除以总观测数 $(n = 60)$ 算出. 例如 $f_1 = \frac{11}{60} = 0.183\ 3$. 累计频数列 F_i 由 f_i 的值递归求和生成. 因此, $F_1 = f_1 = 0.183\ 3, F_2 = F_1 + f_2 = 0.183\ 3 + 0.133\ 3 = 0.316\ 6$.

i	小区间	观测记录符号	观测频数 o_i	相对频数 f_i	累计相对频数 F_i
1	(0, 1]	卌 卌 |	11	0.183 3	0.183 3
2	(1, 2]	卌 |||	8	0.133 3	0.316 6
3	(2, 3]	卌 ||||	9	0.150 0	0.466 6
4	(3, 4]	卌 ||	7	0.116 7	0.583 3
5	(4, 5]	卌 |	6	0.100 0	0.683 3
6	(5, 6]	卌	5	0.083 3	0.766 6
7	(6, 7]	||||	4	0.066 7	0.833 3
8	(7, 8]	||	2	0.033 3	0.866 6
9	(8, 9]	|||	3	0.050 0	0.916 6
10	(9, 10]	|||	3	0.050 0	0.966 6
11	(10, 11]	|	1	0.016 7	0.983 3
12	(11, 12]	|	1	0.016 7	1.000 0
总和			60	1.000 0	

f_i 和 F_i 的值提供了服务时间 t 的 pdf 和 CDF 的等价性. 由于频数直方图给出了连续服务时间的"离散化"形式, 我们可以通过用线段连接得到的点, 从而把得到的 CDF 转换成分段连续的函数. 图 14.6 给出了本例经验 pdf 和 CDF 的图像. 如直方图所示, 各个小区间的中点连线给出了 CDF.

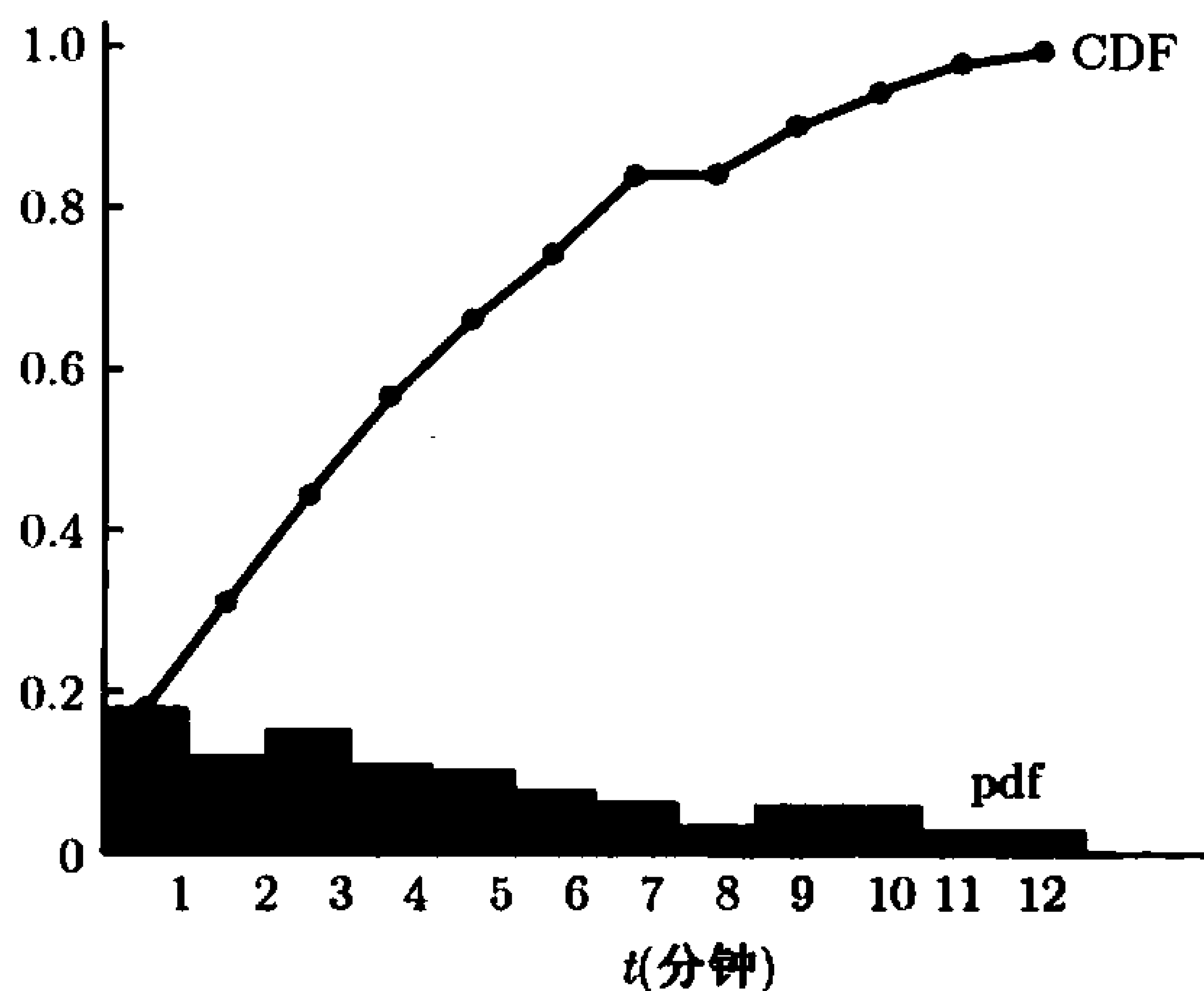


图 14.6 经验分布的分段线性 CDF

现在可以估计出该经验分布的平均值 $\bar{t}$ 和方差 s_t^2 如下. 令 N 为直方图中小区间的个数, 定义 $\bar{t}_i$ 为小区间 i 的中点, 则

$$\bar{t}=\sum_{i=1}^{N} f_i\bar{t}_i$$

$$s_t^2=\sum_{i=1}^{N} f_i(\bar{t}_i-\bar{t})^2$$

用这些公式表示本例, 我们得到

$$\bar{t}=0.183\,3\times 0.5+0.133\times 1.5+\cdots+11.5\times 0.016\,7=3.934 \text{ 分钟}$$

$$s_t^2=0.188\,3\times(0.5-3.934)^2+0.1333\times(1.5-3.934)^2+\cdots$$
$$+0.016\,7\times(11.5-3.934)^2=8.646 \text{ 分钟}^2$$

Excel 程序

用 Excel 的电子表格可以方便地构造直方图. 从菜单栏上选择 **工具** ⇒ **数据分析** ⇒ **直方图**[①], 然后在对话框中输入合适的数据. 但是 Excel 中的*直方图*工具并不把直接产生的频率直方图的平均值和标准差作为输出结果.[②] 因此, Excel 程序 excelMeanVar.xls 是用来计算样本平均值、方差、最大值和最小值的, 并且可以和 Excel 的*直方图*工具同时使用.

图 14.7 在单元格 A8:E19 中存放例 14.5-1 的输入数据. 随着数据输入到电子表格中, 将自动产生样本统计量 (均值、标准差、最小值和最大值).

为了构造出直方图, 首先产生小区间的上限, 并从第 8 行开始输入到 F 列. 在本例中, 单元 F8:F19 用来给出小区间边界, 然后在直方图对话框中输入样本数据的位置和小区间边界 (如图 14.7 下面部分所示).

输入区域: A8:E19

接收区域: F8:F19

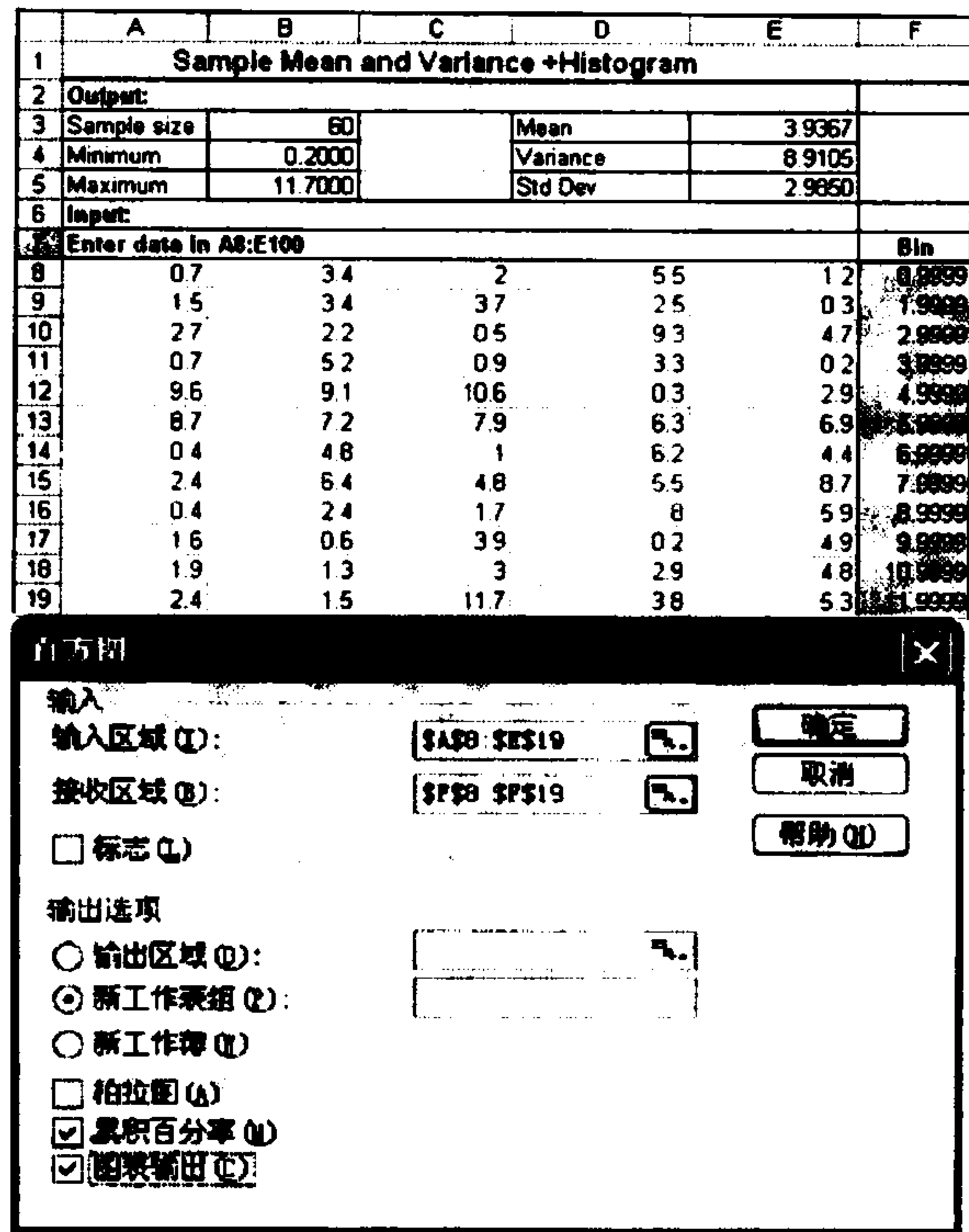
输出选项: 在*累积百分率*和*图表输出*前打勾.

现在, 单击*确定*, 图 14.8 显示了输出结果.

拟合优度检验 拟合优度检验用于评价在确定经验分布时所用的样本是否服从某个特定的理论分布. 可以把经验 CDF 和某个理论分布的 CDF 进行比较, 从而对数据做出初步的评价. 若这两个 CDF 没有明显的差异, 则很有可能这些样本服从这个理论分布. 再用拟合优度检验来进一步确定这一初步的 "预感". 下面的例子给出了详细的方法.

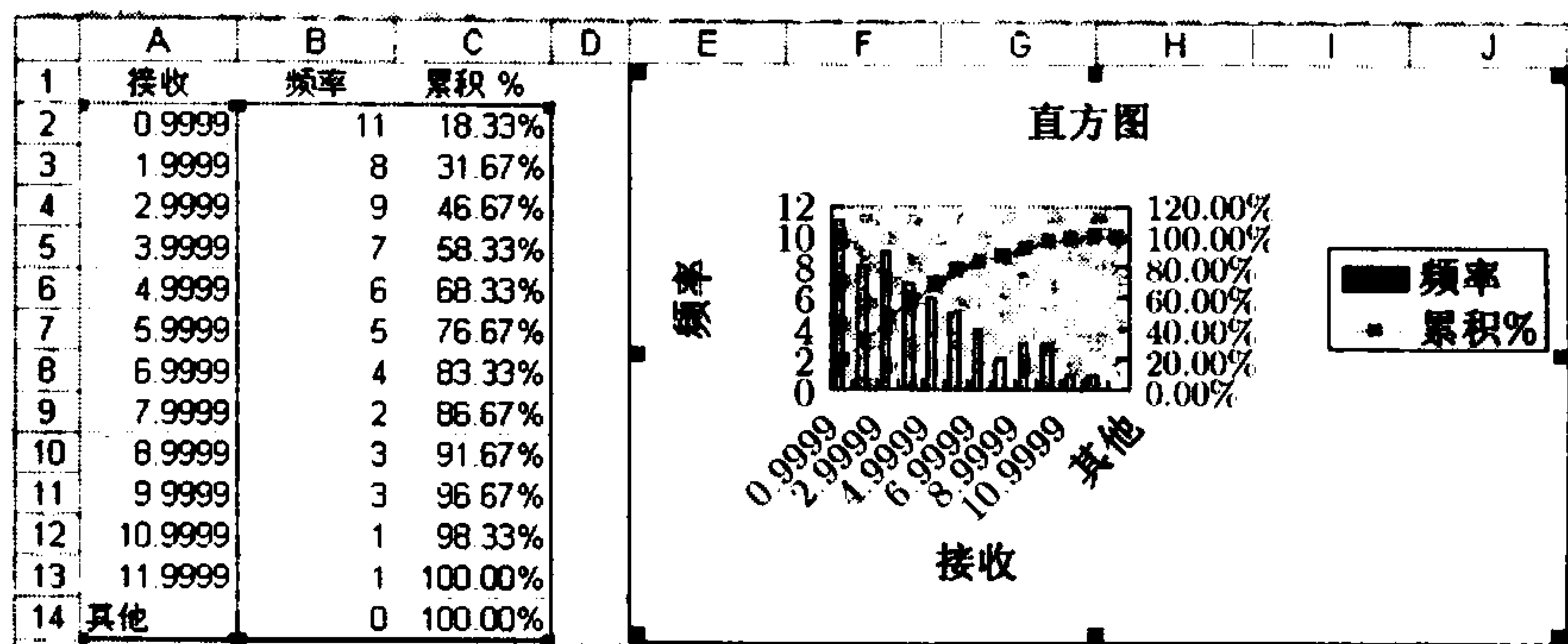
① 如果*数据分析*不出现在*工具*菜单下面, 单击该菜单下的*加载宏*并选择*分析工具库*, 然后再单击*确定*. ——译者注

② Excel 的*数据分析*提供了单独的统计工具 (*描述统计*), 用来计算平均值和方差 (以及其他一些你可能从未使用过的统计功能).



	A	B	C	D	E	F
1	Sample Mean and Variance +Histogram					
2	Output:					
3	Sample size	60		Mean	3.9367	
4	Minimum	0.2000		Variance	8.9105	
5	Maximum	11.7000		Std Dev	2.9850	
6	Input:					
7	Enter data in A8:E100					Bin
8	0.7	3.4	2	5.5	1.2	0.9999
9	1.5	3.4	3.7	2.5	0.3	1.9999
10	2.7	2.2	0.5	9.3	4.7	2.9999
11	0.7	5.2	0.9	3.3	0.2	3.9999
12	9.6	9.1	10.6	0.3	2.9	4.9999
13	8.7	7.2	7.9	6.3	6.9	[illegible]
14	0.4	4.8	1	6.2	4.4	6.9999
15	2.4	6.4	4.8	5.5	8.7	7.9999
16	0.4	2.4	1.7	8	5.9	8.9999
17	1.6	0.6	3.9	0.2	4.9	9.9999
18	1.9	1.3	3	2.9	4.8	10.9999
19	2.4	1.5	11.7	3.8	5.3	11.9999

图 14.7 例 14.5-1 的 Excel 直方图输入数据和对话框



	A	B	C
1	接收	频率	累积 %
2	0.9999	11	18.33%
3	1.9999	8	31.67%
4	2.9999	9	46.67%
5	3.9999	7	58.33%
6	4.9999	6	68.33%
7	5.9999	5	76.67%
8	6.9999	4	83.33%
9	7.9999	2	86.67%
10	8.9999	3	91.67%
11	9.9999	3	96.67%
12	10.9999	1	98.33%
13	11.9999	1	100.00%
14	其他	0	100.00%

图 14.8 例 14.5-1 的 Excel 直方图的输出结果

例 14.5-2

验证例 14.5-1 的数据服从某个猜测的指数分布.

首先要指定定义理论分布的函数. 从例 14.5-1 中, $\bar{t} = 3.934$ 分钟. 因此, 猜想的指数分布的 $\lambda = \frac{1}{3.934} = 0.254\,2$ 次服务/分钟 (见 14.4.3 节), 且相应的 pdf 和

CDF 为

$$f(t)=0.254\ 2\mathrm{e}^{-0.254\ 2t},\quad t>0$$

$$F(T)=\int_0^T f(t)\mathrm{d}t = 1-\mathrm{e}^{-0.254\ 2T},\quad T>0$$

可以用这个 CDF 和 $F(T)$ 来对 $T=0.5,1.5,\cdots,11.5$ 计算理论的 CDF, 然后把例 14.5-1 计算得到的经验值 $F_i, i=1,2,\cdots,12$ 在图形上进行比较. 例如

$$F(0.5)=1-\mathrm{e}^{-0.254\ 2\times 0.5}\approx 0.12$$

图 14.9 给出了比较结果. 这两个图像的粗略比较说明, 这个指数分布的确能够比较好地符合观测数据.

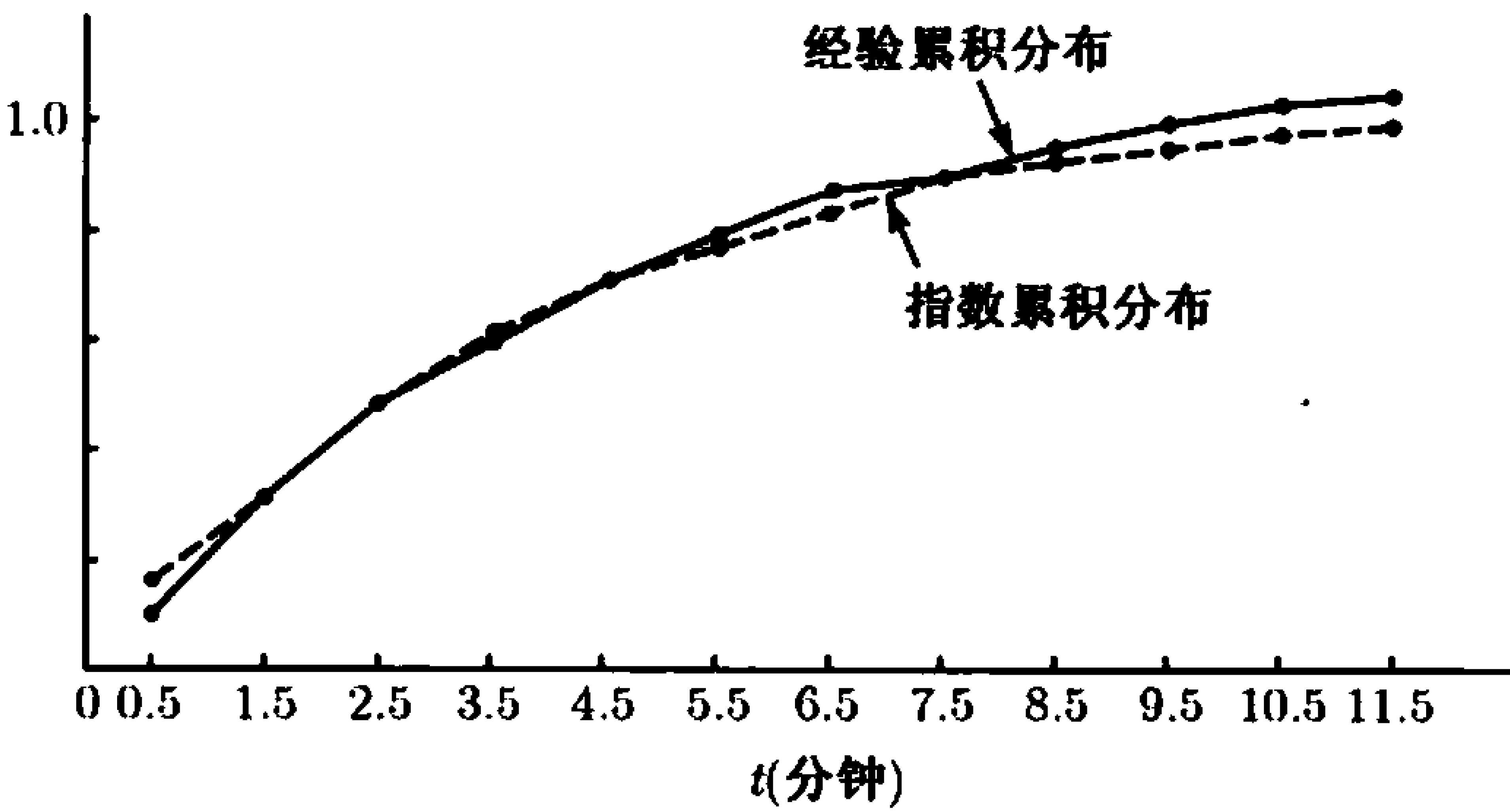


图 14.9　经验 CDF 与理论指数 CDF 的比较

下一步是进行拟合优度检验. 共有两个这样的检验: (1) Kolmogrov-Smirnov 检验; (2) χ^2 检验. 这里只介绍 χ^2 检验法.

χ^2 检验对于直方图的不同小区间, 来度量经验频率和理论频率之间的差异. 特别地, 计算出对应于第 i 个小区间的观测频率 o_i 的理论频率 n_i:

$$n_i=n\int_{I_{i-1}}^{I_i} f(t)\mathrm{d}t = n\left(F(I_i)-F(I_{i-1})\right)=60(\mathrm{e}^{-0.254\ 2I_{i-1}}-\mathrm{e}^{-0.254\ 2I_i})$$

根据给出的直方图小区间 i 的 o_i 和 n_i, 计算经验频率和观测频率之间差异的度量为

$$\chi^2=\sum_{i=1}^{N}\frac{(o_i-n_i)^2}{n_i}$$

当 $N\to\infty$ 时, χ^2 趋向于具有 $N-k-1$ 个自由度的 χ^2 概率密度函数, 其中 k 为从原数据中估计的参数个数, 用来定义理论分布.

零假设说明, 若

$$\chi^2 < \chi^2_{N-k-1,1-\alpha}$$

则接受观测的样本来自该理论分布 $f(x)$ 的结论, 其中 $\chi^2_{N-k-1,1-\alpha}$ 是 $N-k-1$ 自由度和 α 显著性水平的 χ^2 值.

下表给出了检验的计算结果:

i	小区间	观测频数 o_i		理论频数 n_i		$\frac{(o_i-n_i)^2}{n_i}$
1	$(0,1)$	11		13.448		0.453
2	$(1,2)$	8		10.435		0.570
3	$(2,3)$	9		8.095		0.100
4	$(3,4)$	7		6.281		0.083
5	$(4,5)$	6	11	4.873	8.654	0.636
6	$(5,6)$	5		3.781		
7	$(6,7)$	4	9	2.933	6.975	0.588
8	$(7,8)$	2		2.276		
9	$(8,9)$	3		1.766		
10	$(9,10)$	3	5	1.370	6.111	0.202
11	$(10,11)$	1		1.063		
12	$(11,\infty)$	1		3.678		
总和		$n=60$		$n=60$		$\chi^2=2.623$

根据经验, 任何一个小区间内的期望理论频数至少为 5 个. 要满足这一要求, 通常我们把连续的小区间合并起来, 如表中所示, 得到的小区间数目变成 $N=7$. 因为我们从观测数据中只估计一个参数 (即 λ), 所以 χ^2 分布的自由度必须等于 $7-1-1=5$. 如果要求显著性水平 $\alpha=0.05$, 则得到临界值 $\chi^2_{5.05}=11.07$(利用附录 B 中的表 3, 或者在 excelStatTables.xls 的 F8 中输入 5, 在 L8 输入 0.05, 并在 R8 中读出答案). 因为 χ^2 值 (=2.623) 小于临界值, 所以接受假设, 即样本来自于所猜测的指数分布.

习题 14.5A

1. 下表数据表示某个服务设施的到达间隔时间 (单位: 分钟):
   (a) 利用 Excel 程序, 对于小区间宽度 0.5, 1 和 1.5 分钟, 分别作出 3 个直方图.
   (b) 用图像比较经验 CDF 和相应的指数分布的累积分布.
   (c) 检验假设: 根据 95% 的置信水平, 所给出的样本来自指数分布.
   (d) 用来检验零假设, 这 3 个直方图哪一个"最好"?

4.3	3.4	0.9	0.7	5.8	3.4	2.7	7.8
4.4	0.8	4.4	1.9	3.4	3.1	5.1	1.4
0.1	4.1	4.9	4.8	15.9	6.7	2.1	2.3
2.5	3.3	3.8	6.1	2.8	5.9	2.1	2.8
3.4	3.1	0.4	2.7	0.9	2.9	4.5	3.8
6.1	3.4	1.1	4.2	2.9	4.6	7.2	5.1
2.6	0.9	4.9	2.4	4.1	5.1	11.5	2.6
0.1	10.3	4.3	5.1	4.3	1.1	4.1	6.7
2.2	2.9	5.2	8.2	1.1	3.3	2.1	7.3
3.5	3.1	7.9	0.9	5.1	6.2	5.8	1.4
0.5	4.5	6.4	1.2	2.1	10.7	3.2	2.3
3.3	3.3	7.1	6.9	3.1	1.6	2.1	1.9

2. 下表数据表示传送一条信息需要的时间长度 (秒).

25.8	67.3	35.2	36.4	58.7
47.9	94.8	61.3	59.3	93.4
17.8	34.7	56.4	22.1	48.1
48.2	35.8	65.3	30.1	72.5
5.8	70.9	88.9	76.4	17.3
77.4	66.1	23.9	23.8	36.8
5.6	36.4	93.5	36.4	76.7
89.3	39.2	78.7	51.9	63.6
89.5	58.6	12.8	28.6	82.7
38.7	71.3	21.1	35.9	29.2

利用 Excel 程序作出适当的直方图. 已知下列有关理论均匀分布的信息, 按照 95% 的置信水平, 检验假设: 这些数据来自一均匀分布.

(a) 该分布的区间是 0 到 100. (b) 该分布区间是从样本数据估计的.

(c) 分布区间的最大上限为 100, 但最小下限必须从样本数据中估计.

3. 一个自动装置用来在某个繁忙交叉路口计数交通流量. 该装置以连续时间尺度从零开始记录下一辆车到达交叉路口的时间. 下表提供了前 60 辆车的 (累计) 到达时间 (分钟). 用 Excel 程序作出适当的直方图, 然后按 95% 的置信水平检验假设: 这些间隔时间来自某一指数分布.

到达	到达时间 (分钟)	到达	到达时间 (分钟)	到达	到达时间 (分钟)	到达	到达时间 (分钟)
1	5.2	16	67.6	31	132.7	46	227.8
2	6.7	17	69.3	32	142.3	47	233.5
3	9.1	18	78.6	33	145.2	48	239.8
4	12.5	19	86.6	34	154.3	49	243.6
5	18.9	20	91.3	35	155.6	50	250.5
6	22.6	21	97.2	36	166.2	51	255.8
7	27.4	22	97.9	37	169.2	52	256.5
8	29.9	23	111.5	38	169.5	53	256.9
9	35.4	24	116.7	39	172.4	54	270.3
10	35.7	25	117.3	40	175.3	55	275.1
11	44.4	26	118.2	41	180.1	56	277.1
12	47.1	27	124.1	42	188.8	57	278.1
13	47.5	28	127.4	43	201.2	58	283.6
14	49.7	29	127.6	44	218.4	59	299.8
15	67.1	30	127.8	45	219.9	60	300.0

参考文献

Feller, W., *An Introduction to Probability Theory and Its Applications*, 2nd ed., Vols. 1 and 2, Wiley, New York, 1967.

Paulos, J. A., *Innumeracy: Mathematical Illiteracy and Its Consequences*, Hill and Wang, New York, 1988.

Papoulis, A., *Probability and Statistics*, Prentice Hall, Upper Saddle River, NJ, 1990.

Parzen, E., *Modern Probability Theory and Its Applications*, Wiley, New York, 1960.

Ross, S., *Introduction to Probability Models*, 5th ed., Academic Press, New York, 1993.

第 15 章　随机库存模型

本章导读　本章是第 10 章确定性库存模型的续篇, 研究需求为随机性的库存问题. 所建立的模型大致可分为*连续型*和*周期型*的盘点情形. 周期型盘点模型包括单周期和多周期问题, 所提出的求解方法涉及了从在确定性 EOQ 中使用随机项到用动态规划求解更复杂问题的方法. 这里介绍的随机模型表面上似乎"太理论化"而不实用, 但事实上, 第 24 章介绍的一个案例分析正是利用了其中的一个模型帮助戴尔公司解决其库存问题, 从而节约了可观的资金.

本章包含 1 项实际应用问题、4 个例题、1 个 Excel 程序、22 个节后习题和 2 个案例, 其中案例在附录 E 中介绍, AMPL/Excel/Solver/TORA 程序放在 ch15Files 子目录下.

实际应用——戴尔供应链中的库存决策

戴尔公司实行一种直接销售式的经营模式, 在这种模式下, 公司将个人电脑直接销售给美国的客户. 当从客户那里收到一份订单后, 公司将规格要求发送给得克萨斯州奥斯汀的制造厂, 在那里 8 小时内就可以完成计算机的组装、测试和包装. 戴尔公司保持很少的库存, 相反, 它要求其供应商 (通常设在东南亚) 保持所谓的"周转"库存, 它们放在制造厂附近的*周转仓库*(零部件仓库) 那里. 这些周转仓库属于戴尔公司, 租给供应商使用. 然后, 戴尔从周转仓库里"提出"所需要的零部件, 而供应商的责任是及时补充库存, 以满足戴尔的预计需求. 虽然戴尔并不拥有周转仓库中的库存, 但库存的费用通过零部件的价格被间接传递给顾客. 因此, 减少了的库存通过降低产品价格直接使戴尔的顾客受益. 这里所提出的解已经导致每年节约的费用达到近 270 万美元. 第 24 章案例 13 详细介绍了这一案例研究.

15.1　连续盘点模型

本节介绍两个模型: (1) 确定性 EOQ 模型 (10.2.1 节) 的"概率化版本", 它利用一个缓冲库存来代表随机需求; (2) 一个更加纯粹的随机 EOQ 模型, 模型中直接包含了随机需求.

15.1.1　"概率化"的 EOQ 模型

一些实际工作者对确定性的 EOQ 模型 (10.2.1 节) 做了调整, 通过在整个计划期内添加一种固定的缓冲库存, 来近似地反映需求的随机性质. 确定缓冲库存的

大小, 应使得在提前时间(从发出订单到收货之间的间隔时间) 内用完库存的概率不超过某一指定的值.

令

L = 从发出订单到收货之间的提前时间

x_L = 表示在提前时间内需求的随机变量

μ_L = 提前时间期间的平均需求

σ_L = 提前时间期间需求的标准差

B = 缓冲库存规模

α = 提前时间期间用完库存的最大允许概率

该模型的主要假设是, 提前时间期间 L 的需求 x_L 服从平均数为 μ_L、标准差为 σ_L 的正态分布, 即 $N(\mu_L, \sigma_L)$.

图 15.1 表示了缓冲库存 B 和确定性 EOQ 模型参数之间的关系, 这些参数包括提前时间 L、提前时间期间的平均需求 μ_L, 以及 EOQ y^*. 注意到, L 必须等于 10.2.1 节中定义的有效提前时间.

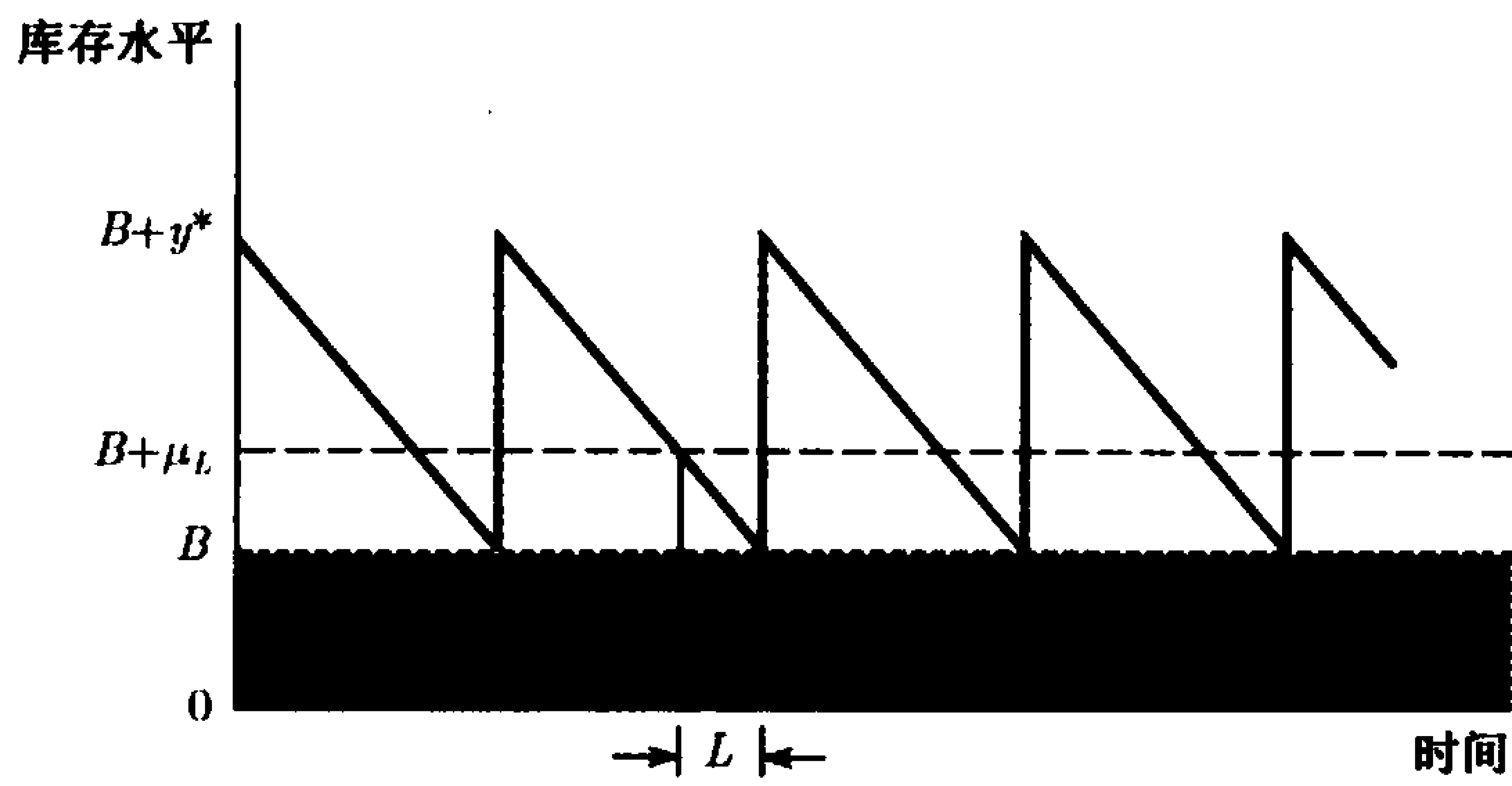


图 15.1 在经典 EOQ 模型上引入缓冲库存

用来确定 B 的概率表达式可以写成

$$P\{x_L \geqslant B + \mu_L\} \leqslant \alpha$$

可以用以下代换, 把 x_L 转换为标准 $N(0,1)$ 的随机变量 (见 14.5.4 节):

$$z = \frac{x_L - \mu_L}{\sigma_L}$$

这样就有

$$P\left\{z \geqslant \frac{B}{\sigma_L}\right\} \leqslant \alpha$$

图 15.2 确定了 K_α(从附录 B 中的标准正态表得到, 或用文件 excelStatTables.xls 求出), 使得

$$P\{z \geqslant K_\alpha\} = \alpha$$

因此, 缓冲库存规模必满足

$$B \geqslant \sigma_L K_\alpha$$

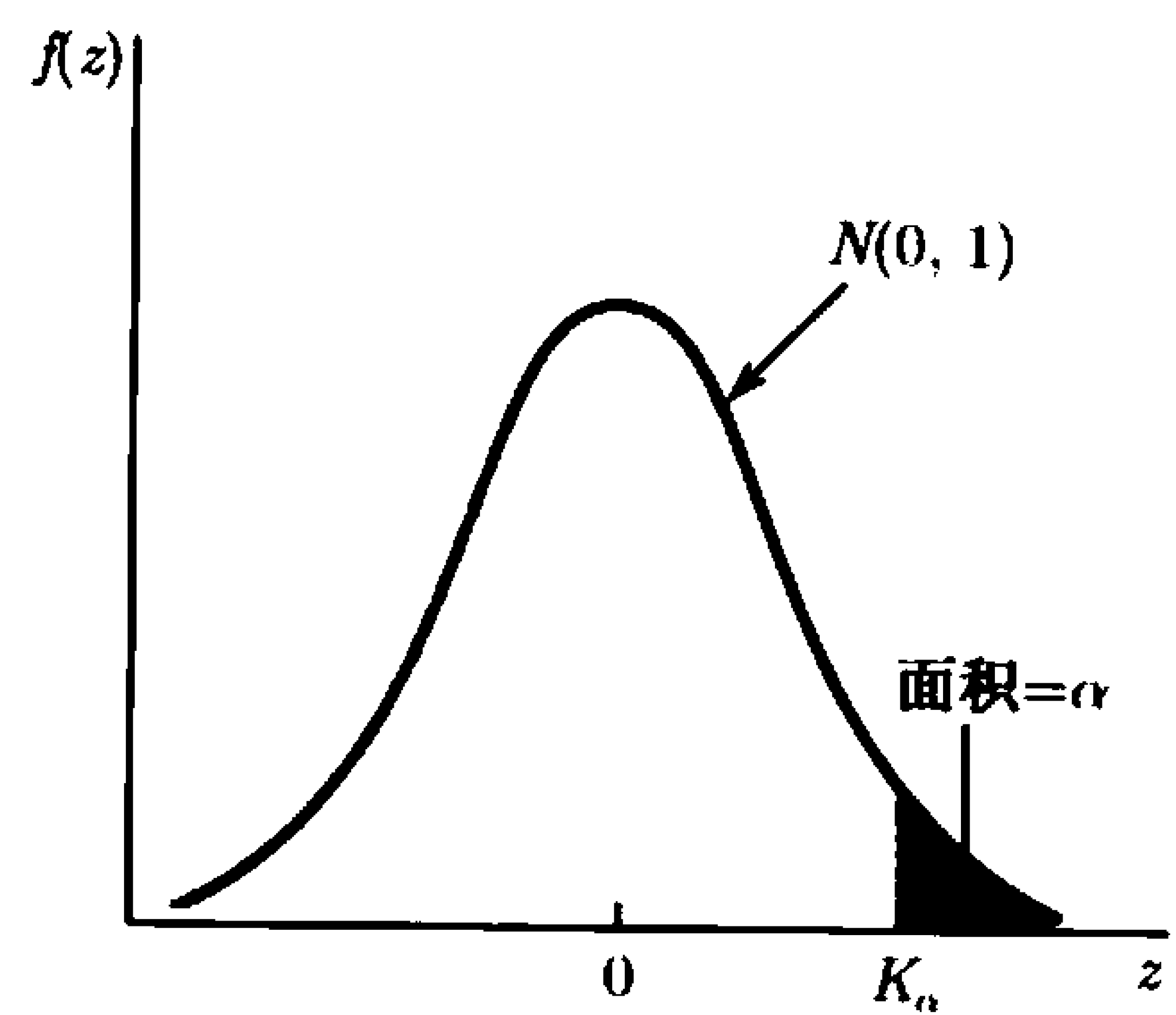


图 15.2 用尽库存的概率 $P\{z \geqslant K_\alpha\} = \alpha$

提前时间 L 期间的需求通常用每单位时间 (如每天或每周) 的某个概率密度函数来描述, 由这个概率密度函数, 就可以确定 L 期间的分布. 设单位时间的需求是正态的, 平均值为 D, 标准差为 σ, 则可以算出提前时间 L 期间的需求平均值 μ_L 和标准差 σ_L 为

$$\mu_L = DL \qquad \sigma_L = \sqrt{\sigma^2 L}$$

计算 σ_L 的公式要求 L 为 (舍入到) 一整数值.

例 15.1-1

在例 10.2-1 确定霓虹灯库存策略的问题中, $EOQ = 1\ 000$ 件. 如果每日需求量为正态的, 均值为 $D = 100$ 只灯管, 标准差 $\sigma = 10$ 只灯管, 即 $N(100, 10)$, 确定缓冲库存规模, 使得用完库存的概率不超过 $\alpha = 0.05$.

由例 10.2-1, 有效提前时间 $L = 2$ 天, 因此

$$\mu_L = DL = 100 \times 2 = 200 \text{ 只}$$
$$\sigma_L = \sqrt{\sigma^2 L} = \sqrt{10^2 \times 2} = 14.14 \text{ 只}$$

给定 $K_{0.05} = 1.645$, 可计算缓冲规模为

$$B \geqslant 14.14 \times 1.645 \approx 23 \text{ 只霓虹灯管}$$

因此, 带有缓冲 B 的最优库存策略要求, 一旦库存水平降至 $223 (= B + \mu_L = 23 + 2 \times 100)$ 只, 则订货 1 000 只.

习题 15.1A

1. 在例 15.1-1 中, 针对下列每种情况, 求出最优库存策略:

*(a) 提前时间 =15 天. (b) 提前时间 =23 天.
(c) 提前时间 =8 天. (d) 提前时间 =10 天.

2. 某音像店卖一种畅销光盘, 该光盘的日需求量 (张数) 接近于正态分布, 平均数为 200 张, 标准差为 20 张. 在店里保存光盘的费用为每天每张 \$0.04, 发出一份新订单的费用是 \$100, 供货有 7 天的提前时间. 假设该店想要把提前时间期间该种光盘脱销的概率限制在 0.02 以内, 确定该店的最优库存策略.

3. 某度假区内的一个礼品店对照相机胶卷的日需求量是正态分布的, 均值为 300 卷, 标准差 5 卷. 店中库存一卷胶卷的费用是 \$0.02, 每次新订货要付的一笔固定订货费用为 \$30. 该店的库存策略要求, 一旦库存水平下降到 80 卷的时候, 就要订货 150 卷, 同时还要随时保持 20 卷的一个缓冲库存量.
   (a) 对于上述库存策略, 确定提前时间期间库存脱销的概率.
   (b) 对于给定的上述数据, 为该店建议一个库存策略, 假设提前时间期间售完胶卷的概率不超过 0.10.

15.1.2 随机 EOQ 模型

没有理由相信, 15.1.1 节的"概率化"EOQ 模型一定能给出最优的库存策略, 因为有关需求的随机性质的信息一开始就没有考虑到, 只是在后来的计算阶段才以一种完全独立的方式"加了进来", 这个事实就足以否定了它的最优性.

为了弥补这一缺陷, 本节给出一个更加精确的模型, 其中模型的表达直接引入了需求的随机性质.

与 15.1.1 节中的情况不同, 这个新的模型允许需求不足, 如图 15.3 所示. 这一策略要求, 一旦库存水平降至 R, 则订货量为 y. 如确定性情形一样, 再订货水平 R 是从订货到收货之间的提前时间的函数. 通过使每单位时间的期望费用 (包括订货费、储存费和缺货费用的总费用) 达到极小, 来求出最优的 y 值和 R 值.

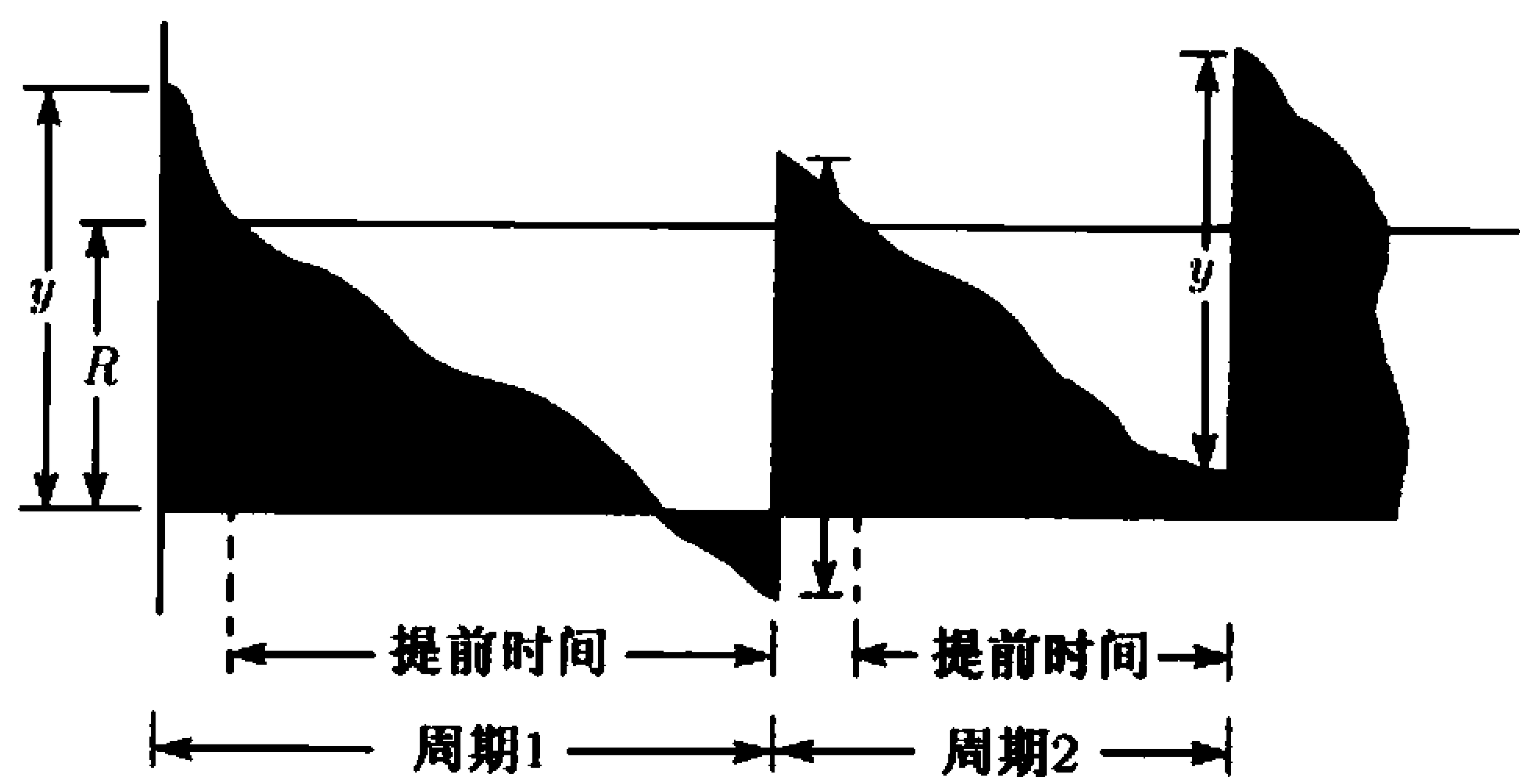


图 15.3 带有缺货的随机库存模型

该模型有 3 个假定.

(1) 提前时间期间未满足的需求被积压起来.

(2) 最多允许一个订单未得以解决.

(3) 提前时间期间的需求分布为时间平稳 (不变) 的.

为建立单位时间的总费用函数, 令

$f(x)$ = 提前时间期间需求 x 的概率密度函数

D = 单位时间的期望需求

h = 单位时间内每库存单位的储存费用

p = 每库存单位的缺货费用

K = 每份订单的订货费用

根据这些定义, 就可以确定费用函数的要素项.

(1) **订货费**. 每单位时间的近似订单数为 $\frac{D}{y}$, 因此每单位时间的订货费用约为 $\frac{KD}{y}$.

(2) **期望储存费用**. 平均库存为

$$I = \frac{(y + \mathrm{E}\{R - x\}) + \mathrm{E}\{R - x\}}{2} = \frac{y}{2} + R - \mathrm{E}\{x\}$$

这个公式基于一个周期的开始期望库存和结束期望库存, 分别为 $y + \mathrm{E}\{R - x\}$ 和 $\mathrm{E}\{R - x\}$. 作为一种近似, 该表达式忽略了 $R - \mathrm{E}\{x\}$ 可能为负的情况. 因此每单位时间的期望储存费用等于 hI.

(3) **期望缺货费用**. 当 $x > R$ 时出现缺货, 因此每个周期的期望缺货量为

$$S = \int_R^\infty (x - R)f(x)\mathrm{d}x$$

因为 p 仅与缺货量成比例, 所以每个订货周期的期望缺货费用为 pS, 且根据每单位时间有 $\frac{D}{y}$ 个订货周期, 所以每单位时间的缺货费用为 $\frac{pDS}{y}$.

得出的每单位时间的总费用函数为

$$TCU(y, R) = \frac{DK}{y} + h\left(\frac{y}{2} + R - \mathrm{E}\{x\}\right) + \frac{pD}{y}\int_R^\infty (x - R)f(x)\mathrm{d}x$$

通过下面的公式求出最优解 y^* 和 R^*:

$$\frac{\partial TCU}{\partial y} = -\left(\frac{DK}{y^2}\right) + \frac{h}{2} - \frac{pDS}{y^2} = 0$$

$$\frac{\partial TCU}{\partial R} = h - \left(\frac{pD}{y}\right)\int_R^\infty f(x)\mathrm{d}x = 0$$

因此我们得到

$$y^* = \sqrt{\frac{2D(K + pS)}{h}} \tag{1}$$

$$\int_{R^*}^\infty f(x)\mathrm{d}x = \frac{hy^*}{pD} \tag{2}$$

因为从 (1) 和 (2) 式得不到闭形式的 y^* 和 R^*, 所以我们用了 Hadley and Witin (1963, pp. 169-174) 给出的一种数值算法来寻找最优解. 假如存在可行解的话, 该算法经过有限步迭代终止.

对于 $R=0$, 上面的 (1) 和 (2) 式给出

$$\hat{y}=\sqrt{\frac{2D(K+pE\{x\})}{h}} \qquad \tilde{y}=\frac{PD}{h}$$

如果 $\tilde{y}\geqslant\hat{y}$, 则 y 和 R 存在唯一的最优值. 按照求解过程, 当 $S=0$ 时, 得到 y^* 的最小值为 $\sqrt{\frac{2KD}{h}}$.

算法的计算步骤如下.

第 0 步 利用初始解 $y_1=y^*=\sqrt{\frac{2KD}{h}}$, 并令 $R_0=0$. 设定 $i=1$, 转到第 i 步.

第 i 步 用 y_i 从公式 (2) 得到 R_i. 如 $R_i\approx R_{i-1}$, 停止, 最优解为 $y^*=y_i$, $R^*=R_i$; 否则, 在公式 (1) 中代入 R_i 计算 y_i. 令 $i=i+1$, 重复第 i 步.

例 15.1-2

Electro 公司在其生产过程中要使用树脂, 每月用量 1 000 加仑. Electro 公司为每次新订货要支付 \$100 的订货费, 每加仑树脂每月的储存费用是 \$2, 而且每加仑的缺货费用为 \$10. 历史数据表明, 提前时间期间的需求服从 $(0,100)$ 加仑之间的均匀分布. 请为 Electro 公司确定最优的订货策略.

利用模型变量符号, 我们有

$$D=1\ 000\ \text{加仑/月} \qquad K=\text{每次订货 \$100}$$
$$h=\text{每加仑每月 \$2} \qquad p=\text{每加仑 \$10}$$
$$f(x)=\frac{1}{100}, 0\leqslant x\leqslant 100 \qquad E\{x\}=50\ \text{加仑}$$

首先, 我们需要检查该问题是否有可行解. 利用 $\hat{y}$ 和 $\tilde{y}$ 的公式, 我们得到

$$\hat{y}=\sqrt{\frac{2\times 1\ 000(100+10\times 50)}{2}}=774.6\ \text{加仑}$$
$$\tilde{y}=\frac{10\times 1\ 000}{2}=5\ 000\ \text{加仑}$$

因为 $\tilde{y}\geqslant\hat{y}$, 故存在唯一的最优解 y^* 和 R^*.

S 的公式计算如下:

$$S=\int_R^{100}(x-R)\frac{1}{100}\mathrm{d}x=\frac{R^2}{200}-R+50$$

用公式 (1) 和 (2) 中的 S, 我们得到

$$y_i=\sqrt{\frac{2\times 1\ 000(100+10S)}{2}}=\sqrt{100\ 000+10\ 000S}\ \text{加仑} \tag{3}$$
$$\int_R^{100}\frac{1}{100}\mathrm{d}x=\frac{2y_i}{10\times 1\ 000}$$

由最后一个公式可得

$$R_i = 100 - \frac{y_i}{50} \tag{4}$$

现在用公式 (3) 和 (4) 求最优解.

迭代 1

$$y_1 = \sqrt{\frac{2KD}{h}} = \sqrt{\frac{2 \times 1\ 000 \times 100}{2}} = 316.23 \text{ 加仑}$$

$$R_1 = 100 - \frac{316.23}{50} = 93.68 \text{ 加仑}$$

迭代 2

$$S = \frac{R_1^2}{200} - R_1 + 50 = 0.199\ 71 \text{ 加仑}$$

$$y_2 = \sqrt{100\ 000 + 10\ 000 \times 0.199\ 71} = 319.37 \text{ 加仑}$$

因此,

$$R_2 = 100 - \frac{319.39}{50} = 93.612 \text{ 加仑}$$

迭代 3

$$S = \frac{R_2^2}{200} - R_2 + 50 = 0.203\ 99 \text{ 加仑}$$

$$y_3 = \sqrt{100\ 000 + 10\ 000 \times 0.203\ 99} = 319.44 \text{ 加仑}$$

因此,

$$R_3 = 100 - \frac{319.44}{50} = 93.611 \text{ 加仑}$$

由于 $y_3 \approx y_2$, $R_3 \approx R_2$, 最优解为 $R^* \approx 93.611$ 加仑, $y^* \approx 319.44$ 加仑. 可以使用文件 excelContRev.xls 来求任何所需精度的最优解. 本例最优库存策略为, 当库存水平下降到 94 加仑时, 订货量大约为 320 加仑.

习题 15.1B

1. 用例 15.1-2 给出的数据, 求下列各量.
   (a) 每月近似的订单数目.
   (b) 期望月度订货费.
   (c) 每月期望储存费用.
   (d) 每月期望缺货费用.
   (e) 提前时间期间用完库存的概率.

*2. 解例 15.1-2, 假定提前时间期间的需求为 0~50 加仑的均匀分布.

*3. 在例 15.1-2 中, 假设提前时间期间的需求是 40~60 加仑的均匀分布. 请计算最优解并与例 15.1-2 得出的解进行比较, 解释比较结果 (提示: 这两个问题的 $E\{x\}$ 相同, 但本问题的方差较小.)

4. 找出例 15.1-2 的最优解, 假定提前时间期间的需求服从 $N(100, 2)$. 设 $D = 10\ 000$ 加仑/月, $h =$ 每月每加仑 \$2, $p =$ 每加仑 \$4, $K = \$20$.

15.2 单周期模型

单种货品的库存模型适用于当一件货品只订货一次来满足该周期的需求, 比如, 时装到季节末就会过时. 本节介绍两个模型, 分别表示没有订货费和带有订货费的两种情况.

建模所用的变量符号有

$$
\begin{aligned}
K &= \text{每份订单的订货费用}\\
h &= \text{该周期期间每库存单位的储存费用}\\
p &= \text{周期期间每缺货单位的惩罚费用}\\
D &= \text{随机变量, 表示该周期期间的需求}\\
f(D) &= \text{周期期间需求的概率密度函数}\\
y &= \text{订货量}\\
x &= \text{新订货之前的现有库存}
\end{aligned}
$$

这一模型确定 y 的最优值, 使得期望储存费用和缺货费用之和达到最小. 给定了最优的 $y(=y^*)$ 后, 库存策略是: 如 $x < y$, 订货 $y^* - x$; 否则不订货.

15.2.1 没有订货费的模型 (报摊模型)

这个模型在文献中又称为报摊模型 (最早的经典称谓是报童模型), 因为它针对的是像报纸一类短寿命的货品.

模型的假设包括:

(1) 需求在周期开始随着收到订货后立即产生;

(2) 没有订货费.

图 15.4 显示了需求 D 满足以后的库存位置. 如果 $D < y$, 则在周期期间保持量 $y - D$; 否则, 如果 $D > y$, 则出现缺货量 $D - y$.

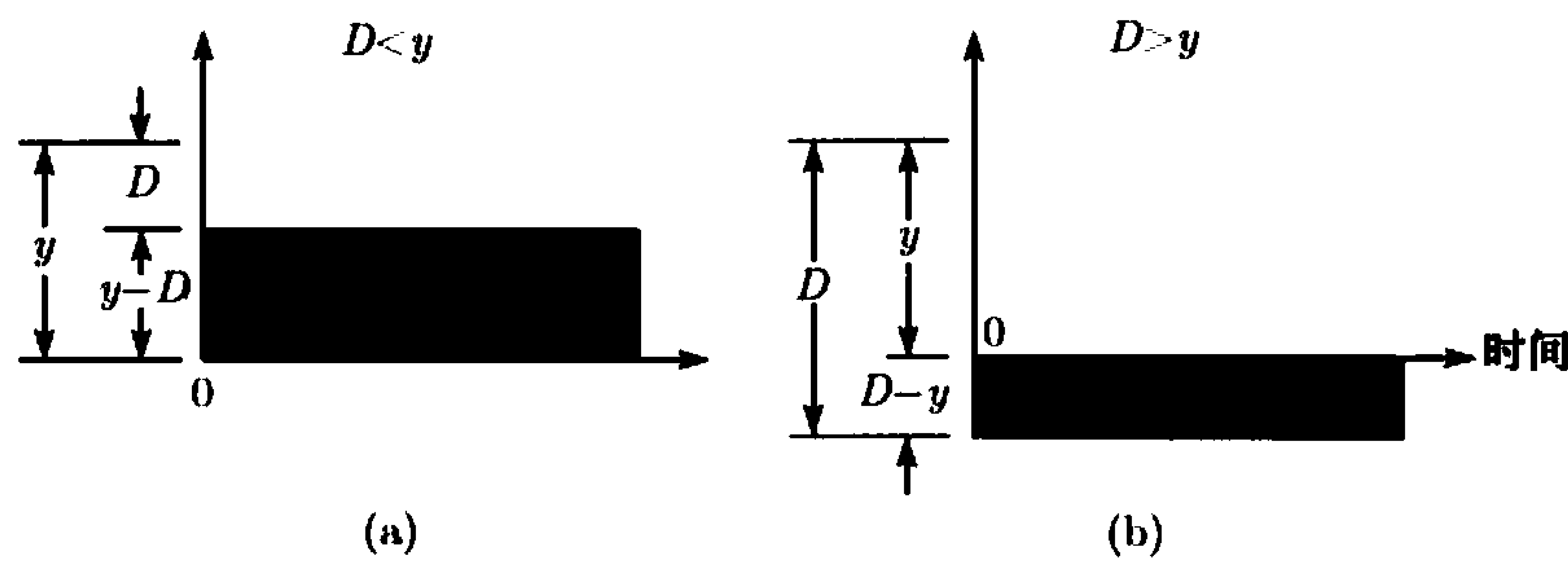


图 15.4 单周期模型中的储存和缺货库存

该周期的期望费用 $E\{C(y)\}$ 的表达式为

$$E\{C(y)\} = h\int_0^y (y-D)f(D)\mathrm{d}D + p\int_y^\infty (D-y)f(D)\mathrm{d}D$$

函数 $E\{C(y)\}$ 对 y 是凸的, 故可证明它有唯一的最小值. 求 $E\{C(y)\}$ 关于 y 的一阶导数并令其等于 0, 我们得到

$$h\int_0^y f(D)\mathrm{d}D - p\int_y^\infty f(D)\mathrm{d}D = 0$$

或

$$hP\{D \leqslant y\} - p(1 - P\{D \leqslant y\}) = 0$$

或

$$P\{D \leqslant y^*\} = \frac{p}{p+h}$$

前面的推导假定需求 D 是连续的. 如果 D 是离散的话, 则 $f(D)$ 只在离散点有定义, 并且相应的费用函数是

$$E\{C(y)\} = h\sum_{D=0}^{y} (y-D)f(D) + p\sum_{D=y+1}^{\infty} (D-y)f(D)$$

最优性必要条件是

$$E\{C(y-1)\} \geqslant E\{C(y)\} \text{ 和 } E\{C(y+1)\} \geqslant E\{C(y)\}$$

由于 $E\{C(y)\}$ 是凸函数, 因此这些条件也是充分的. 经过若干代数运算, 运用这些条件可以得到下面计算 y^* 的不等式:

$$P\{D \leqslant y^* - 1\} \leqslant \frac{p}{p+h} \leqslant P\{D \leqslant y^*\}$$

例 15.2-1

某个报亭主想要确定每天早晨必须上货的 *USA Now* 报纸的数量. 该报亭主付 30 美分进一份报纸, 然后以 75 美分一份卖出. 报纸的销售通常在早上 7:00 到 8:00 之间, 晚上剩下的报纸只能按 5 美分一份回收. 假定每天的需求如下, 该报亭主每天早晨该进多少份报纸呢?

(1) 服从平均值 300 份, 标准差 20 份的正态分布.

(2) $f(D)$ 为离散的概率密度函数, 定义如下:

D	200	220	300	320	340
$f(D)$	0.1	0.2	0.4	0.2	0.1

该问题并没有明确定义储存和缺货惩罚费用. 问题的数据表明, 对每份没有卖出的报纸, 报亭主要损失 $30-5=25$ 美分, 但若上货不够的话每份报纸的惩罚费用为 $75-30=45$ 美分. 这样, 按照库存问题的参数, 我们有 $h=$ 每份每天 25 美分, $p=$ 每份每天 45 美分.

首先, 我们确定临界比为

$$\frac{p}{p+h}=\frac{45}{45+25}=0.643$$

情形 1 需求 D 是 $N(300,20)$. 可以使用 excelStatTables.xls 来计算最优订货量, 在 F15 中输入 300, 在 G15 中输入 20, 在 L15 中输入 0.643, 就能在单元格 R15 得到所需要的答案 307.33 份报纸. 也可以使用附录 B 的标准正态表查出该值. 定义

$$z=\frac{D-300}{20}$$

然后从表中查

$$\mathrm{P}\{z\leqslant 0.366\}\approx 0.643$$

或

$$\frac{y^*-300}{20}=0.366$$

因此, $y^*=307.3$. 最优订货量大约为 308 份.

情形 2 需求 D 服从某离散概率密度函数 $f(D)$. 首先确定离散分布函数 $\mathrm{P}\{D\leqslant y\}$ 为

y	200	220	300	320	340
$\mathrm{P}\{D\leqslant y\}$	0.1	0.3	0.7	0.9	1.0

对计算得到的临界比 0.643, 我们有

$$\mathrm{P}(D\leqslant 220)\leqslant 0.643\leqslant \mathrm{P}(D\leqslant 300)$$

只能得到 $y^*=300$ 份.

习题 15.2A

1. 对于单周期模型, 证明离散型需求的最优订货量可以从下式求出:

$$\mathrm{P}\{D\leqslant y^*-1\}\leqslant\frac{p}{p+h}\leqslant \mathrm{P}\{D\leqslant y^*\}$$

2. 某种货品在单个周期间的需求在该周期开始瞬间出现, 相应的概率密度函数服从 10~15 个单位的均匀分布. 由于很难估计费用参数, 订货量的确定要使得出现剩余或短缺的概率不超过 0.1. 能同时满足这两个条件吗?

*3. 某单周期库存问题中单位储存费用为 \$1. 假如订货量是 4 件, 找出最优性条件要求的单位惩罚费用的允许范围. 假设需求在周期开始时瞬时出现, 并且需求的概率密度函数如下表所示.

D	0	1	2	3	4	5	6	7	8
$f(D)$	0.05	0.1	0.1	0.2	0.25	0.15	0.05	0.05	0.05

4. 某大学书店为教授们提供课堂教材复印业务. Yataha 教授教大学一年级的课程, 听课学生数服从 200~250 名的均匀分布. 书店复印一份教材的成本为 \$10, 并以 \$25 一份的价格卖给学生. 学生们只在学期开始时购买这些教材, 如果卖不完就只能回收. 同时, 一旦书店缺货, 就不能再多复印了, 学生们只能到其他地方去买. 假如书店想要获得最大的利润, 它应该印多少份教材呢?

5. QuickStop 便利店每天早晨 6 点钟为顾客供应咖啡和多那圈 (一种油炸圈饼). 该便利店以 7 美分/张的价格购进多那圈, 并在早晨 8 点前以 25 美分/张卖给顾客. 过了 8 点钟后, 这些多那圈只能卖 5 美分/张. 在 6 点到 8 点之间购买多那圈的顾客人数为 30~50 人的均匀分布. 每位顾客通常买 3 个多那圈加一份咖啡. 要获得最大利润, QuickStop 便利店每天早晨大约要备多少打 (每打 12 张) 多那圈呢?

*6. Colony 商场正在进一批大衣, 准备冬季来临时销售. Colony 商场花 \$50 买一件大衣, 以 \$110 一件卖出. 当冬季结束时, Colony 以 \$55 一件促销这些大衣. 冬季期间对该大衣的需求量为 20 件以上, 但不超过 30 件, 为等概率的. 由于冬季不长, 单位储存费用可忽略不计. 另外, Colony 的经理认为短缺并不会导致任何惩罚性费用. 请确定最优订货量, 使得该商场的收益达到最大. 可以采用连续的近似方法.

7. 对于单周期模型, 假设货品的消费在周期期间是均匀分布的 (而不是在周期开始时瞬时出现). 建立相应的费用模型, 并求出最优订货量.

8. 求解例 15.2-1, 假定周期期间的需求是连续的和均匀的, 且需求的概率密度函数在 0~100 是均匀的 (提示: 利用第 7 题的结果).

15.2.2 带有订货费的模型 (s-S 策略)

本节模型和 15.2.1 节模型的区别是带有一项订货费用 K. 用同样的符号, 每周期的总期望费用为

$$\begin{aligned}\mathrm{E}\{\overline{C}(y)\} &= K + \mathrm{E}\{C(y)\} \\ &= K + h\int_0^y (y-D)f(D)\mathrm{d}D + p\int_y^{\infty}(D-y)f(D)\mathrm{d}D\end{aligned}$$

如 15.2.1 节所推导的, 最优值 y^* 必须满足

$$\mathrm{P}\{y \leqslant y^*\} = \frac{p}{p+h}$$

因为 K 是常数, $\mathrm{E}\{\overline{C}(y)\}$ 的最小值也必在 y^* 取得.

在图 15.5 中, $S = y^*$, 并且 $s(< S)$ 的值由以下方程求出:

$$\mathrm{E}\{C(s)\} = \mathrm{E}\{\overline{C}(S)\} = K + \mathrm{E}\{C(S)\}, s < S$$

该方程产生另外一个值 $s_1(> S)$, 这个值我们没有用.

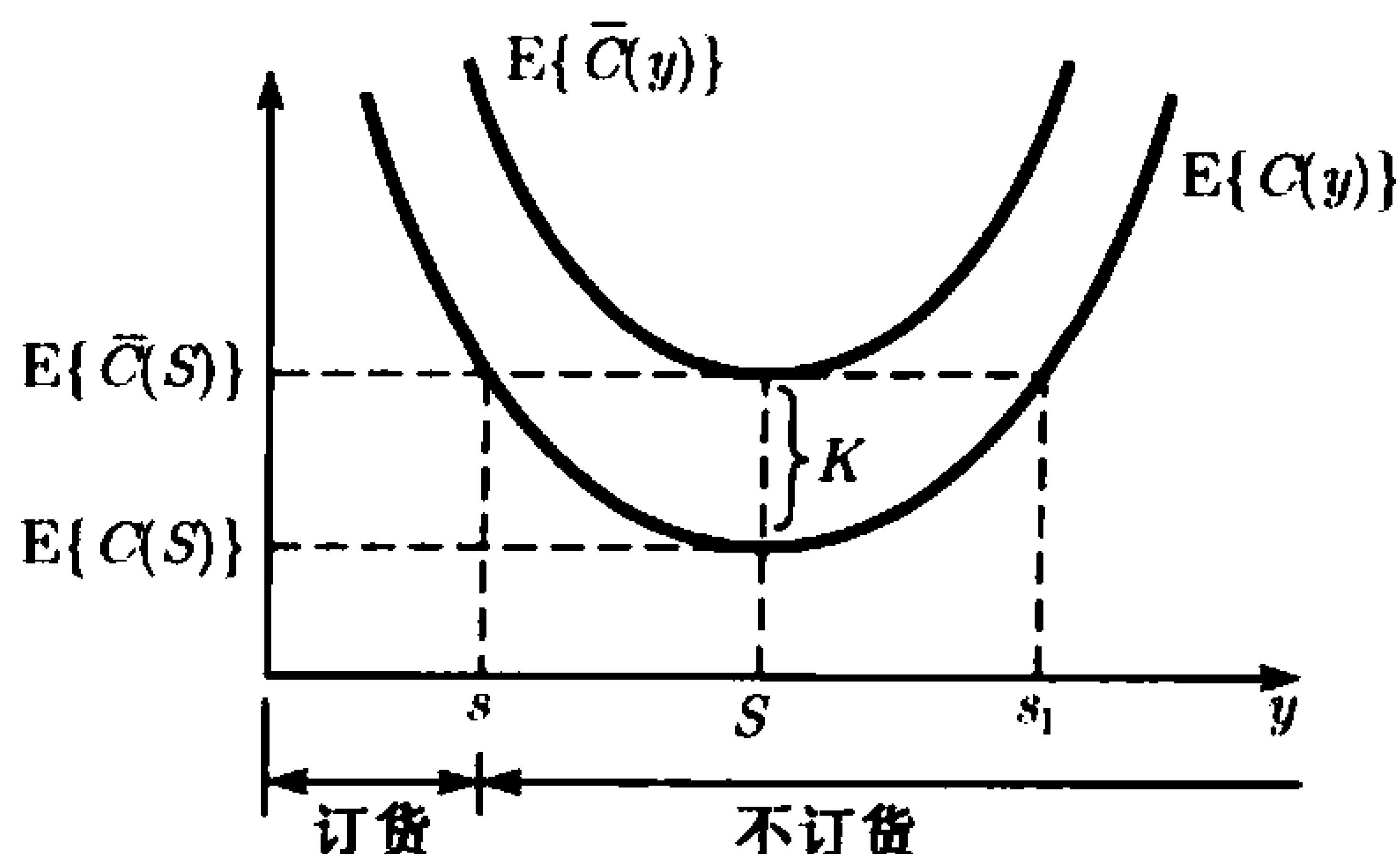


图 15.5 带有订货费的单周期库存模型中的 (s-S) 最优订货策略

如果订货之前的现有存量为 x 件, 那么应该订货多少件呢? 我们针对下面 3 个条件对这个问题进行研究:

(1) $x < s$ (2) $s \leqslant x \leqslant S$ (3) $x > S$

情形 1 ($x<s$) 因为已经有 x 件存量, 等价费用由 $\mathrm{E}\{C(x)\}$ 给出. 如果另外订货 $y - x$ $(y > x)$, 则按照 y, 相应的费用就是 $\mathrm{E}\{\overline{C}(y)\}$, 其中包括订货费用 K. 由图 15.5, 我们有

$$\min_{y>x} \mathrm{E}\{\overline{C}(y)\} = \mathrm{E}(\overline{C}(S)) < \mathrm{E}\{C(x)\}$$

因此, 这一情形下的最优订货量为 $S - x$ 件.

情形 2 ($s \leqslant x \leqslant S$) 由图 15.5, 我们有

$$\mathrm{E}\{C(x)\} \leqslant \min_{y>x} \mathrm{E}\{\overline{C}(y)\} = \mathrm{E}(\overline{C}(S))$$

因此, 这一情形下的订货是不利的, 故 $y^* = x$.

情形 3 ($x>S$) 由图 15.5, 对 $y > x$ 我们有

$$\mathrm{E}\{C(x)\} < \mathrm{E}\{\overline{C}(y)\}$$

这个条件表示这一情形下的订货是不利的, 即 $y^* = x$.

最优库存策略, 通常也称为 s-S 策略, 归纳如下:

如果 $x < s$, 则订货 $S - x$;

如果 $x \geqslant s$, 则不订货.

由于相应的费用函数为凸, s-S 策略的最优性是有保证的.

例 15.2-2

在一单周期内某种货品的日需求在周期开始时瞬时出现, 需求的概率密度函数为 0~10 个单位的均匀分布. 周期内货品的单位储存费用为 \$0.50, 用完库存的单位惩罚费用为 \$4.50. 每次订货要支付一笔固定的订货费 \$25. 请确定该货品的最优库存量.

为求出 y^*, 考虑

$$\frac{p}{p+h}=\frac{4.5}{4.5+0.5}=0.9$$

以及

$$\mathrm{P}\{D\leqslant y^*\}=\int_0^{y^*}\frac{1}{10}\mathrm{d}D=\frac{y^*}{10}$$

因此, 有 $S=y^*=9$.

期望费用函数由下式给出:

$$\begin{aligned}\mathrm{E}\{C(y)\}&=0.5\int_0^{y}\frac{1}{10}(y-D)\mathrm{d}D+4.5\int_y^{10}\frac{1}{10}(D-y)\mathrm{d}D\\&=0.25y^2-4.5y+22.5\end{aligned}$$

为得到 s, 解

$$\mathrm{E}\{C(s)\}=K+\mathrm{E}\{C(S)\}$$

从而得到

$$0.25s^2-4.5s+22.5=25+0.25S^2-4.5S+22.5$$

由于 $S=9$, 上述方程化简为

$$s^2-18s-19=0$$

得到这一方程的解为 $s=-1$ 或 $s=19$, 舍去 $s>S$ 的值. 因为剩下的值是负的 $(=-1)$, s 没有可行值 (图 15.6). 当费用函数是"扁平的", 或者当与模型其他费用相比订货费非常高的时候, 通常就会出现这样的结论.

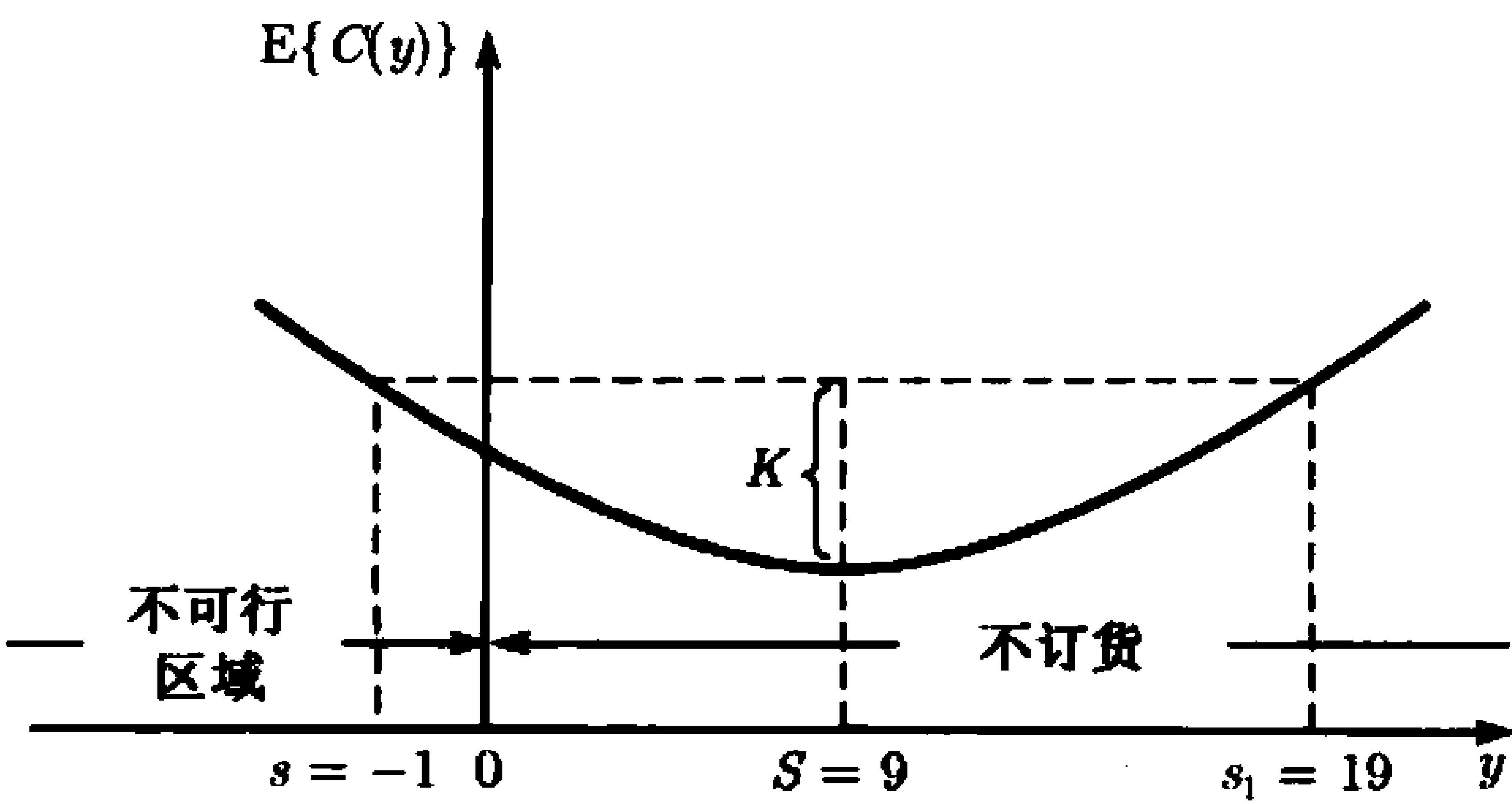


图 15.6 应用于例 15.2-2 的 s-S 策略

习题 15.2B

*1. 求例 15.2-2 的最优库存策略, 假设订货费是 \$5.

2. 在 15.2.1 节的单周期模型中, 假设模型是求最大利润, 而且出现订货费 K. 给定 r 为单位销售价格, 利用 15.2.1 节中的信息建立期望利润的表达式, 并求出最优订货量. 对于 $r = \$3$, $c = \$2$, $p = \$4$, $h = \$1$, $K = \$10$, 用数值方法求解该问题. 需求的概率密度函数服从 0~10 的均匀分布.

3. 求解习题 15.2A 的第 5 题, 假定每次多那圈供货收取 \$10 的固定费用.

15.3 多周期模型

本节介绍一种在没有订货费假定下的多周期模型. 此外, 该模型还允许需求的积压 (backlog), 并没有供货延迟. 模型进一步假设, 任何周期的需求 D 由一平稳的概率密度函数 $f(D)$ 描述.

多周期模型考虑货币的折扣值. 如果 $\alpha(<1)$ 是每周期的折扣系数, 则从现在开始的 n 个周期的一个货币量 $\$A$ 有现值 $\$\alpha^n A$.

设该库存问题包含 n 个周期, 未满足的需求只能往后积压一个周期. 定义

$F_i(x_i) =$ 周期 $i, i+1, \cdots, n$ 的最大期望利润, 其中 x_i 是周期 i 订货之前的存货量.

利用 15.2 节的符号, 并假定 c 和 r 分别为单位费用和收益, 下面建立该库存问题的动态规划模型 (见第 22 章):

$$
\begin{aligned}
F_i(x_i) = \max_{y_i \geqslant x_i} \Bigg\{ & -c(y_i - x_i) + \int_0^{y_i} [rD - h(y_i - D)] f(D) \mathrm{d}D \\
& + \int_{y_i}^{\infty} [ry_i + \alpha r(D - y_i) - p(D - y_i)] f(D) \mathrm{d}D \\
& + \alpha \int_0^{\infty} F_{i+1}(y_i - D) f(D) \mathrm{d}D \Bigg\}, i = 1, 2, \cdots, n
\end{aligned}
$$

其中 $F_{n+1}(y_n - D) = 0$. 由于未满足的需求被积压, 所以 x_i 的值可能小于零. 第 2 个积分中包含量 $\alpha r(D - y_i)$, 这是因为 $(D - y_i)$ 是周期 i 未能满足的需求, 必须在周期 $i+1$ 中补货.

可以用递归方法求解该问题. 对于周期数是无穷的情形, 递归公式化简成

$$
\begin{aligned}
F(x) = \max_{y \geqslant x} \Bigg\{ & -c(y - x) + \int_0^{y} [rD - h(y - D)] f(D) \mathrm{d}D \\
& + \int_{y}^{\infty} [ry + \alpha r(D - y) - p(D - y)] f(D) \mathrm{d}D \\
& + \alpha \int_0^{\infty} F(y - D) f(D) \mathrm{d}D \Bigg\}
\end{aligned}
$$

其中 x 和 y 分别为收货前后每个周期的库存水平.

y 的最优值可根据下列必要条件求得, 因为期望收益函数 $F(x)$ 是凹的, 该条件也是充分条件.

$$\begin{aligned}\frac{\partial(\cdot)}{\partial y} &= -c - h\int_0^y f(D)\mathrm{d}D + \int_y^\infty [(1-\alpha)r + p]f(D)\mathrm{d}D \\ &\quad + \alpha\int_0^\infty \frac{\partial F(y-D)}{\partial y} f(D)\mathrm{d}D = 0\end{aligned}$$

$\frac{\partial F(y-D)}{\partial y}$ 值的计算如下. 若在下一个周期开始还有存货 $\beta(>0)$, 下一周期的利润将会增加 $c\beta$, 因为如此, 订货量要少得多, 这意味着

$$\frac{\partial F(y-D)}{\partial y} = c$$

因此必要条件变成

$$-c - h\int_0^y f(D)\mathrm{d}D + [(1-\alpha)r + p]\left(1 - \int_0^y f(D)\mathrm{d}D\right) + \alpha c\int_0^\infty f(D)\mathrm{d}D = 0$$

最优库存水平 y^* 可由下式求出:

$$\int_0^{y^*} f(D)\mathrm{d}D = \frac{p + (1-\alpha)(r-c)}{p + h + (1-\alpha)r}$$

给定库存水平 x, 则每个周期的最优库存策略为:

如果 $x < y^*$, 则订货 $y^* - x$; 如果 $x \geqslant y^*$, 则不订货.

习题 15.3A

1. 考虑一个两周期随机库存模型, 其中需求被积压, 供货延迟为 0. 每个周期的需求概率密度函数是 0 到 10 之间均匀的, 各费用参数为:

   单位销售价格 =$2　　每月单位储存费用 =$0.10
   单位采购价格 =$1　　每月缺货单位惩罚费用 =$3
   折扣系数 =0.8

   假定周期 1 的初始库存为零, 求出这两个阶段的最优库存策略.

*2. 在无限长时间期间的库存模型中, 每周期需求的概率密度函数为

$$f(D) = 0.08D, 0 \leqslant D \leqslant 5$$

   单位费用参数有

   单位销售价格 =$10　　每月单位储存费用 =$1
   单位采购价格 =$8　　每月缺货单位惩罚费用 =$10
   折扣系数 =0.9

假定没有供货延迟且未满足的需求可积压, 求最优库存策略.

3. 考虑无限长时间期间零延迟及积压需求的库存问题. 根据费用极小化求出最优库存策略. 假定

$$z \text{ 个单位的储存费} = hz^2 \qquad z \text{ 个单位的缺货惩罚费用} = px^2$$

对于 $h = p$ 的特殊情况, 说明最优解与需求的概率密度函数无关.

参 考 文 献

Cohen, R. and F. Dunford, "Forecasting for Inventory Control: An Example of When 'Simple' Means 'Better'," *Interfaces*, Vol. 16, No.6, pp.95-99, 1986.

Hadley, G., and T. Whitin, *Analysis of Inventory Systems*, Prentice Hall, Upper Saddle River, NJ, 1963.

Nahmias, S., *Production and Operations Analysis*, 5th ed., Irwin, Homewood, IL, 2005.

Silver, E., D. Pyke, and R. Peterson, *Decision Systems for Inventory Management and Production Control*, 3rd ed., Wiley, New York, 1998.

Tersine, R., *Principles of Inventory and Materials Management*, North Holland, New York, 1982.

Zipken, P., *Foundations of Inventory Management*, McGraw-Hill, Boston, 2000.

第16章　仿真模型

本章导读　仿真是用来检测实际系统的又一种非常好的工具，它通过对一个系统中随机行为的计算机模拟来估计这个系统的性能度量. 基本上，仿真都是把一种运行状况看作是一个服务设备上的等待线. 通过实际跟踪顾客在服务设备中的移动，就可以收集到一些相关的统计信息 (例如，等待时间和队列长度). 仿真工作首先要建立可以用来收集所需数据的计算机模型. 目前已经有一些计算机语言可用于处理这些繁琐的计算工作.

将模型运行一个任意的时间周期，然后把得到的结果看作是完全真实的，这是仿真中常见的误区. 事实上，仿真的结果是随着程序运行的时间而改变的 (有时改变是非常彻底的). 因此，仿真模型用于处理这样的统计实验，它的输出结果需要用适当的统计检验来解释. 在学习本章的过程中，请特别关注仿真实验中的某些特性，包括：(1) 在各种概率分布的抽样中，介于 (0,1) 之间的随机数的重要作用；(2) 用于收集观测资料来满足真实统计实验的基础假设的一些特殊方法.

学习本章之前，需要了解一些基本的概率论和统计学的知识. 另外排队论的知识对本章的学习也有很大帮助.

本章包括 10 个例题、2 个 Excel 模板和 44 个节后习题. AMPL/Excel/Solver /TORA 程序放在文件夹 ch16Files 中.

16.1　蒙特卡罗仿真

蒙特卡罗 (Monte Carlo) 方法是仿真中最常用的方法，这种方法根据随机抽样来估计模型中一些随机的和确定的参数. 蒙特卡罗方法应用的实例很多，包括计算多重积分、估计常数 π ($\cong 3.14159$)，以及求矩阵的逆.

本节将使用一个例题来说明蒙特卡罗方法. 这个例题用于强调仿真实验中的统计性质.

例 16.1-1

利用蒙特卡罗抽样来估计由下面式子定义的圆的面积：

$$(x-1)^2+(y-2)^2=25$$

圆的半径是 $r=5$ cm，圆心是 $(x,y)=(1,2)$.

在估计这个圆的面积之前，首先将圆放在一个边长恰好等于圆直径的正方形内，如图 16.1 所示. 这个正方形的顶点坐标也可以计算出来.

估计圆面积需要假定正方形内部所有的点都是等可能出现的. 下面从正方形内随机取 n 个点, 如果其中 m 个落在圆内, 那么

$$(\text{圆面积的估计}) \cong \frac{m}{n}(\text{正方形的面积}) = \frac{m}{n}(10 \times 10)$$

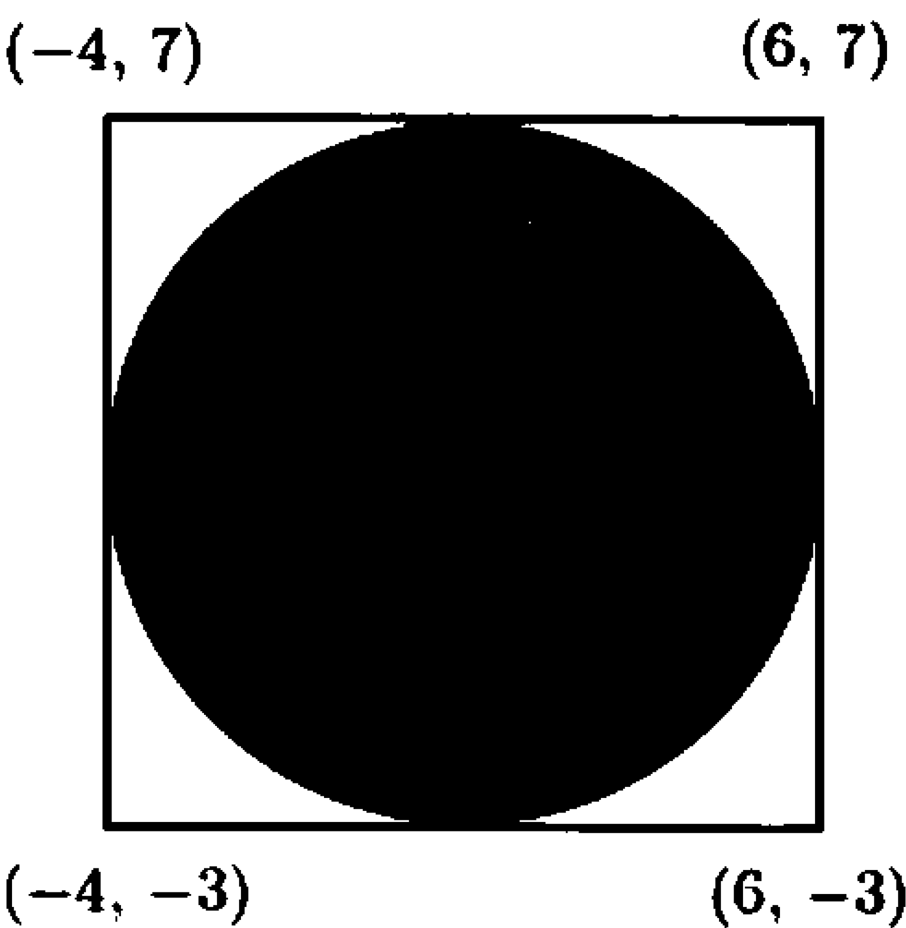


图 16.1 使用蒙特卡罗方法估计圆的面积

为了保证正方形内的所有点出现的概率相等, 我们用下面的均匀分布来表示正方形内一个点的 x 坐标和 y 坐标:

$$f_1(x) = \frac{1}{10}, \quad -4 \leqslant x \leqslant 6$$

$$f_2(y) = \frac{1}{10}, \quad -3 \leqslant y \leqslant 7$$

根据分布 $f_1(x)$ 和 $f_2(y)$ 得到的样本点 (x, y) 可以保证正方形内所有的点可以被等可能地选中.

利用区间 $(0, 1)$ 上的独立均匀分布随机数来确定样本点 (x, y). 表 16.1 提供了一个本例题计算所用到的随机数清单. 根据仿真的目的, 可以采用特殊的算术运算来生成所需要的 0-1 随机数, 16.4 节将详细介绍.

表 16.1 一个简单的 0-1 随机数表

0.058 9	0.352 9	0.586 9	0.345 5	0.790 0	0.630 7
0.673 3	0.364 6	0.128 1	0.487 1	0.769 8	0.234 6
0.479 9	0.767 6	0.286 7	0.811 1	0.287 1	0.422 0
0.948 6	0.893 1	0.821 6	0.891 2	0.953 4	0.699 1
0.613 9	0.391 9	0.826 1	0.429 1	0.139 4	0.974 5
0.593 3	0.787 6	0.386 6	0.230 2	0.902 5	0.342 8
0.934 1	0.519 9	0.712 5	0.595 4	0.160 5	0.603 7
0.178 2	0.635 8	0.210 8	0.542 3	0.356 7	0.256 9
0.347 3	0.747 2	0.357 5	0.420 8	0.307 0	0.054 6
0.564 4	0.895 4	0.292 6	0.697 5	0.551 3	0.030 5

给出一对 0-1 随机数 R_1 和 R_2, 那么图 16.1 正方形中的一个随机点 (x, y) 的坐标可以表示为:

$$x = -4 + [6 - (-4)]R_1 = -4 + 10R_1$$

$$y = -3 + [7 - (-3)]R_2 = -3 + 10R_2$$

为说明该方法的使用, 考虑 $R_1 = 0.058\,9$ 和 $R_2 = 0.673\,3$, 那么

$$x = -4 + 10R_1 = -4 + 10 \times 0.058\,9 = -3.411$$

$$y = -3 + 10R_2 = -3 + 10 \times 0.673\,3 = 3.733$$

这个点落在圆内, 因为

$$(-3.411-1)^2+(3.733-2)^2=22.46<25$$

重复上面的过程 n 次, 记录下落在圆内的点的数目是 m, 那么可以估计出圆的面积为 $\frac{100m}{n}$.

评注 为了增强圆面积估计的可靠性, 我们采用在一般统计实验中所使用的方法:

(1) 增大样本数量. (2) 采用重复实验的方法.

在例题 16.1-1 的讨论中出现了两个有关仿真实验的问题:

(1) 样本数量 n 应该选择多大? 即 n 应该取多少?

(2) 实验需要重复多少次? 即 N 应该取多少?

在统计理论中存在关于如何确定 n 和 N 的一些公式, 它们依赖于具体仿真实验的性质和所需要的置信水平. 然而, 对于任何的统计实验, 有一条黄金准则, 那就是 n 和 N 的取值越大, 仿真的结果就会越可靠. 最后, 样本数量的大小也取决于进行仿真实验的相关费用的大小. 一般来说, 如果选择的样本数量可以得到一个相对较小的标准差, 那么这个样本数量就是足够的.

由于实验输出中的随机性, 因此有必要将结果表示为一个置信区间. 引入 $\overline{A}$ 和 s 分别表示 N 次重复实验的均值和方差, 那么对于置信水平 α, 真实的面积 A 的置信区间是:

$$\overline{A}-\frac{s}{\sqrt{N}}t_{\frac{\alpha}{2},N-1}\leqslant A\leqslant\overline{A}+\frac{s}{\sqrt{N}}t_{\frac{\alpha}{2},N-1}$$

给定了置信水平 α 和自由度 $N-1$, 那么根据 t 分布表可以确定参数 $t_{\frac{\alpha}{2},N-1}$(参见附录 B 的 t 表或者使用 excelStatTables.xls). 要注意的是, N 表示重复实验的次数, 它与样本数量的大小 n 不同.

Excel 程序

由于例题 16.1-1 中相应于每一次实验的计算量比较大, 我们使用 Excel 模板 ExcelCircle.xls(包含 VBA 宏) 来测试样本数量和实验次数对于面积估计的准确度的影响. 输入的数据包括圆的半径、圆心 (cx, cy)、样本数量 n, 以及重复实验的次数 N. 在单元格 D4 中的输入 Steps 说明在同一次执行中允许使用多个样本量. 例如, 当 $n=30\,000$, Steps= 3, 模板最后输出是分别对于 $n=30\,000, 60\,000, 90\,000$ 的计算结果. 每点击一次命令按钮 Press [illegible], 就会得到一个新的面积估计, 这是因为每次 Excel 将更新随机生成的数用于一次新的计算.

图 16.2 给出了对于 5 次重复实验和样本空间大小分别是 30 000, 60 000, 90 000 时的输出结果. 精确的面积是 78.54 cm², 对于 3 个不同的样本量, 采用蒙特卡罗方法得到的面积的平均估计是从 $\overline{A}=78.533$ 到 $\overline{A}=78.490$ cm². 我们还

注意到标准差从 $n = 30\ 000$ 时的 $s = 0.308$ 减小到 $n = 90\ 000$ 时的 $s = 0.191$, 可以得出随着样本量的增大, 解的精确度也会提高.

1	Monte Carlo Estimation of the Area of a Circle			
2	Input data			
3	Nbr. Replications, N =	5		
4	Sample size, n =	30,000	Steps =	3
5	Radius, r =	5		
6	Center, cx =	1		
7	Center, cy =	2		
8	Output results			
9	Exact area =	78 540		
10				
11	Monte Carlo Calculations			
12		n=30000	n=60000	n=90000
13	Replication 1	78.207	78.555	78.483
14	Replication 2	78 673	78.752	78.581
15	Replication 3	78 300	78 288	78 281
16	Replication 4	78 503	78 347	78 343
17	Replication 5	78 983	78 775	78 760
18				
19	Mean =	78 533	78 543	78 490
20	Std. Deviation =	0 308	0 225	0 191
21				
22	95% lower conf limit =	78 151	78 263	78 253
23	95% upper conf limit =	78 915	78 823	78 727

图 16.2 估计圆面积的蒙特卡罗方法的 Excel 输出 (文件 excelCircle.xls)

通过这个实验, 我们感兴趣的是根据最大的样本量输入 (例如 $n = 90\ 000$) 得到置信区间. 给定 $N = 5$, $\overline{A} = 78.490\ \text{cm}^2$, $s = 0.191$, $t_{0.025,4} = 2.776$, 那么可以到 95% 的置信区间是 $78.25 \leqslant A \leqslant 78.73$. 一般而言, 为保证置信区间估计的准确度, N 的取值至少应为 5.

习题 16.1A

1. 在例题 16.1-1 中, 使用表 16.1 中前两列的 $(0, 1)$ 随机数来估计圆的面积. (为简便起见, 依次选择 R_1 然后选择 R_2) 最后再将这个估计与图 16.2 给出的估计进行比较, 有何异同?
2. 假定一个圆的方程是

$$(x - 3)^2 + (y + 2)^2 = 16$$

   (a) 首先定义出相应的分布 $f(x)$ 和 $f(y)$, 然后说明如何根据 $(0, 1)$ 上的随机数对 (R_1, R_2), 来确定一个样本点 (x, y).
   (b) 给定 $n = 100\ 000$ 和 $N = 10$, 使用 ExcelCircle.xls 来估计圆的面积和相应 95% 的置信区间.
3. 利用蒙特卡罗抽样方法来估计图 16.3 给出的湖面的面积. 使用表 16.1 中前两列的 $(0, 1)$ 随机数.
4. 下面考虑 Jan 和 Jim 两个人掷硬币的游戏, 假设硬币是均匀的. 如果结果是正面朝上, 那么 Jan 给 Jim \$10, 否则, Jim 给 Jan \$10.

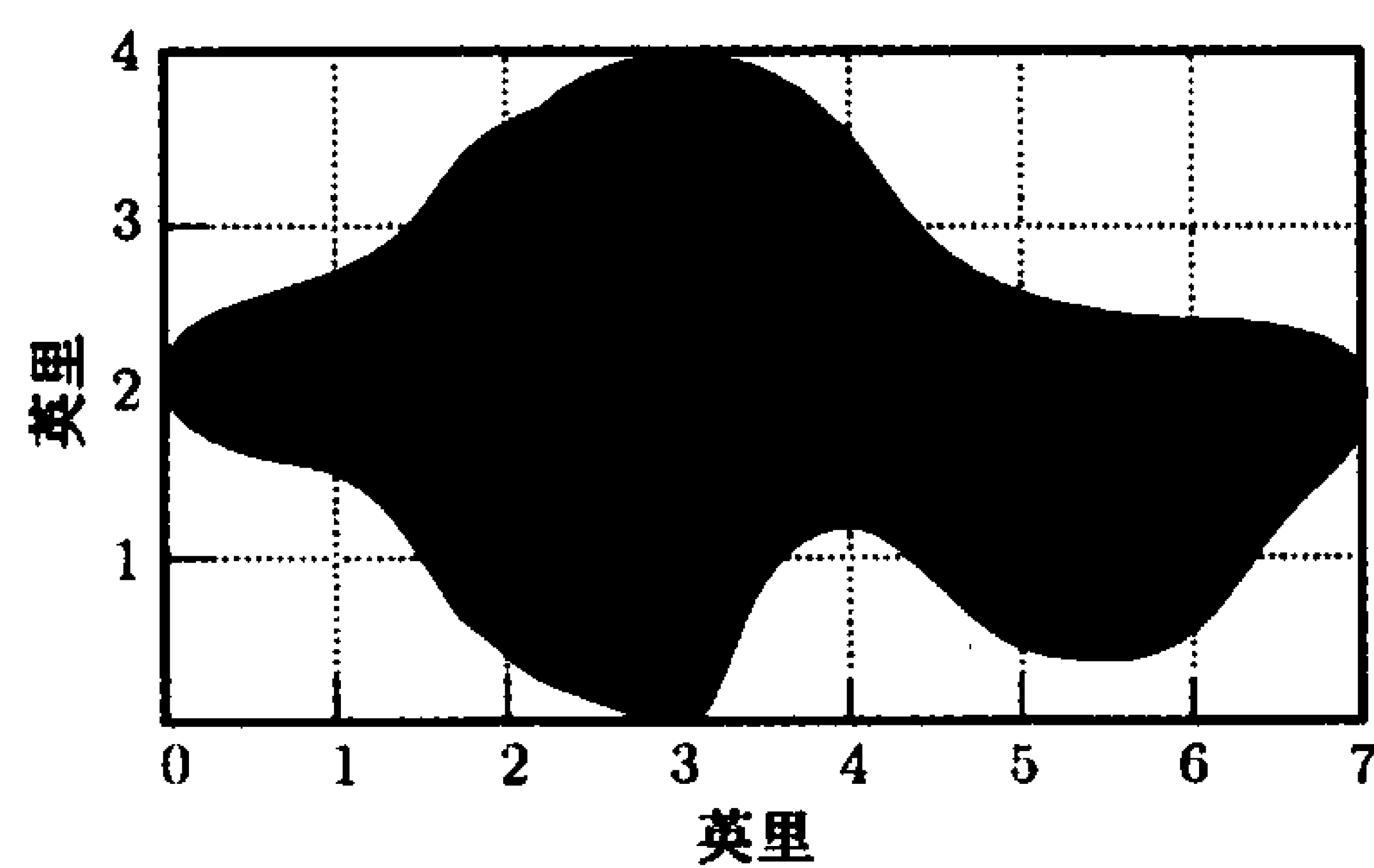


图 16.3 16.1A 第 3 题中的湖面地图

*(a) 如何将这个游戏模拟为一个蒙特卡罗实验?

(b) 每次实验掷 10 次, 重复 5 次实验. 使用表 16.1 中前 5 列的 (0, 1) 随机数, 每一列对应一次重复实验.

(c) 确定 Jan 获胜的 95% 的置信区间.

(d) 比较 (c) 中得到的置信区间与 Jan 理论上赢的期望.

5. 考虑下面的定积分:

$$\int_0^1 x^2 \mathrm{d}x$$

(a) 利用蒙特卡罗方法来估计这个积分的值.

(b) 使用表 16.1 中前 4 列的值来估计这个积分, 其中每次实验的样本数量为 5, 重复 4 次实验. 计算出 95% 的置信区间, 并将这个值与积分的精确值进行比较.

6. 对于下面的掷骰子游戏, 模拟出 5 次输或者赢: 玩家掷两枚均匀骰子, 如果得到的数的和是 7 或者 11, 那么这个玩家将得到 \$10; 否则, 玩家要记录下来这次掷出的和 (点数), 然后继续掷骰子, 直到掷出的点数与记录下来的点数一致, 这种情况下玩家得到 \$10. 如果在这个过程中, 首先得到了 7, 那么玩家将输掉 \$10.

7. 接到订单后的提前时间可以是 1 天或者 2 天, 而且两种情况是等概率的. 每天的需求分别依概率 0.2, 0.7, 0.1 取值为 0, 1, 2. 使用表 16.1 中的随机数 (从第 1 列开始) 来估计需求和提前时间的联合分布. 从这个联合分布, 来估计在提前时间内需求的概率分布函数. (提示: 假定在提前时间内需求取 0~4 的离散值.)

8. 考虑蒲丰 (Buffon) 投针问题的实验. 平面上有多条平行线, 相邻两条平行线相距 D cm, 一根长度为 d cm($d < D$) 的针随机地投放到平面上. 实验的目标是确定这根针不接触以及不与这些平行线相交的概率. 定义

$$h = \text{从针的中心到平行线的垂直距离}$$

$$\theta = \text{针相对于平行线的倾斜角度}$$

(a) 证明针接触或者穿过一条线仅当

$$h \leqslant \frac{d}{2}\sin\theta,\ 0 \leqslant h \leqslant \frac{D}{2},\ 0 \leqslant \theta \leqslant \pi$$

(b) 设计蒙特卡罗实验, 并且给出概率的估计.

(c) 使用 Excel 来得到概率的估计, 每次实验 10 个样本, 重复 4 次. 确定这个估计的 95% 置信区间. 假定 $D = 20$ cm 和 $d = 10$ cm.

(d) 证明: 可以通过下面的式子得到理论上的概率

$$p = \frac{2d}{\pi D}$$

(e) 利用 (c) 的结果和 (d) 的式子来估计 π.

16.2 仿真的类型

如今仿真的执行过程绝大多数是基于蒙特卡罗方法的抽样思想. 所不同的是它把实际系统的行为作为一个*时间的函数*. 存在着两种不同的仿真模型.

(1) **连续模型** 处理行为随着时间*连续*变化的系统. 这些模型通常采用差分 - 微分方程来描述系统中不同元素之间的相互作用. 一个典型的例子就是研究世界人口的动态变化.

(2) **离散模型** 主要是处理排队问题, 目标是确定像平均等待时间和队列长度这样的一些指标, 并且只有当一个客户进入或者离开系统时, 这些指标才会发生变化. 这种事件的变化只是在一些特殊的时间点上发生 (例如一个顾客到达或者离开的时刻), 因此对这种模型取名为**离散事件仿真** (discrete event simulation).

本章给出了离散事件仿真的基础知识, 包括仿真模型的组成部分、仿真统计的收集和模拟实验统计方面的性质. 本章还强调了计算机和仿真语言在执行仿真模型中的作用.

习题 16.2A

1. 将下列情形进行分类, 或者是离散型或者是连续型 (或者是两者的组合). 在每一个案例中, 指出建立仿真模型的目标.

   *(a) 订单随机到达一个仓库. 但是订单不能立即从现有库存中得到满足, 必须等待新货到达.

   (b) 世界人口正在受到自然资源的可用量、粮食生产、环境条件、教育水平、医疗保健和资本投资的影响.

   (c) 货物放在一个自动化仓库接收间的托盘上. 将托盘放在一个较低的传送带上, 并通过电梯将托盘提升到一个较高的传送带上, 最后把托盘传输到走廊, 通过行走式起重机, 从传送带上拿起托盘, 放进储存箱.

2. 解释你为什么会同意或不同意下面的说法: "大多数的离散事件模型都可以描述为某种形式的排队系统, 其中这个系统包括产生客户的*源点*、顾客需要等候的*队列*, 以及为顾客提供服务的*设备*".

16.3 离散事件仿真的要素

本节介绍事件仿真的概念, 并且说明仿真系统的统计数据是如何收集的.

16.3.1 事件的一般定义

所有的离散事件仿真都直接或者间接地描述了一个排队系统, 其中包括顾客到达、在队列中等待 (如果需要的话), 以及得到服务, 最后离开系统. 一般地, 任何一个离散事件模型都由一个相关的队列的网络组成.

鉴于离散事件模型实际上都是一些队列系统的合成, 那么只有在顾客到达或者离开设备时, 才进行仿真统计量 (例如, 队长和服务设备的状态) 的收集. 这就意味着有两个重要的事件控制着仿真模型: 到达和离开. 也就是在这两个时刻我们需要去检查系统, 而在其他时刻没有影响系统统计信息的变化发生.

例 16.3-1

Metalco 加工车间接到了加工两种类型的工件: 常规的和紧急的. 所有的工作都要在两台衔接的机器上加工, 两台机器之间的缓冲区足够长. 总是假定紧急的工件比常规的工件有非抢占的优先加工的权利. 确定这个过程中的事件.

这个过程中包含了两个前后相连的队列, 分别对应两台机器. 首先, 一般会倾向于确定这个过程中的下列事件:

A11: 一个常规的工件到达第一台机器

A21: 一个紧急的工件到达第一台机器

D11: 一个常规的工件离开第一台机器

D21: 一个紧急的工件离开第一台机器

A12: 一个常规的工件到达第二台机器

A22: 一个紧急的工件到达第二台机器

D12: 一个常规的工件离开第二台机器

D22: 一个紧急的工件离开第二台机器

而实际上, 我们只要考虑两个事件: 一个新的工件到达车间和一个加工完毕的工件离开机器. 首先注意到事件 D11 和 A12 实际上是同一个事件. 同样, 事件 D21 和 A22 也一样. 再者, 在离散仿真中, 可以用同一个事件 (到达或者离开) 来描述两种类型的工件, 然后将这个事件标记一个**属性** (attribute) 来表示这个工件的类型是常规的或者紧急的. (可以将这种情形下的这个属性看作一个*个人身份证号码*, 而实际上也是这么回事.) 根据上面的解释, 这个模型中的事件就可以简化为两个: (1) 一个到达事件 A(到达车间); (2) 一个离开事件 D(从机器上离开). 与离开事件相关联的行为依赖于它们所在的机器.

在定义了仿真模型中的基本事件以后，我们来看这个模型是怎么执行的. 图 16.4 用图示化的方法直观地表示出了仿真时间轴上事件发生的情况. 在一个事件的所有行为都执行完毕以后，仿真系统用 "跳" 的方法到达下一个事件，实质上，仿真都是在事件发生的时刻执行的.

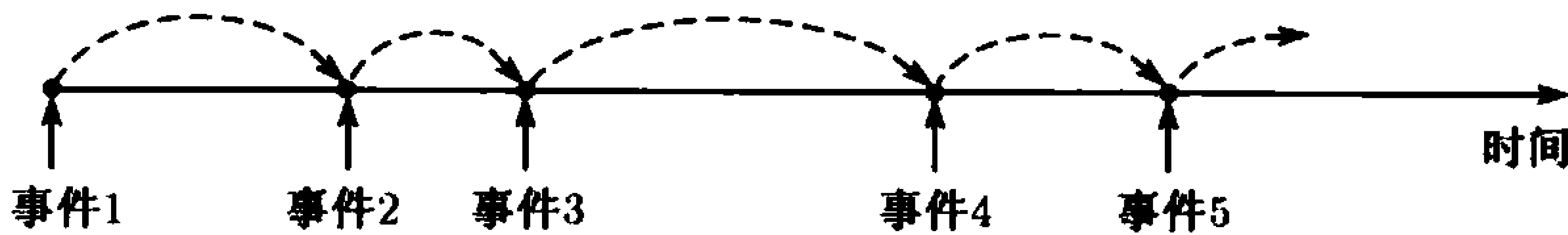


图 16.4 时间轴上仿真事件发生的例子

那么仿真系统是如何确定事件发生的时刻的呢？到达的事件根据它们到达的时间间隔 (两个连续事件到达的时间间隔) 区分开，离开的事件是在设备上服务时间的函数. 所有的这些时间可以是确定的 (如每隔 5 分钟有一列火车到达车站)，或者也可以是随机的 (如银行里顾客到达的时间是随机的). 如果事件之间的时间是确定的，那么它们发生的时间也是确定的. 如果它们之间的时间是随机的，那么我们采用一种特定的方法从一个相应的概率分布中取样. 这一点在 16.3.2 节中将进一步讨论.

习题 16.3A

1. 确定模拟以下情况所需要的离散事件：两种工件从两个不同的来源到达. 这两种工件在同一台机器上加工，优先考虑从第一来源到达的工件.
2. 工件以恒定的到达率到达一个旋转式传送系统. 围绕旋转式传送带布置了 3 个等距的服务站. 如果当一个工件到达服务站时服务器是空闲的，那么工件将从传送带上取下来进行加工；否则，工件将继续在传送带上旋转直到有服务站空闲. 加工过的工件放置在一个额外的储存区域. 确定在这个仿真中的离散事件.
3. 汽车到达了一个双车道的银行停车处，每条车道最多可以容纳 4 辆车. 如果两个车道都满了，新到的车就要到其他地方的银行办理业务. 在任何时间，如果一个车道上的车辆数比另一车道上的车辆数至少多两辆，那么在较长车道上的最后一辆车将会移到较短的车道的后面. 每一个工作日中银行开放停车处的时间是从上午 8:00 到下午 3:00. 定义这种情形下的离散事件.

*4. Elmdale 小学的自助餐厅为所有学生提供了单一通道、固定菜单的午餐. 小朋友每 30 秒时间到达分发窗口，然后每人经过 18 秒得到自己的午餐. 将前 5 个小朋友到达-离开的事件描述在时间轴上.

16.3.2 从概率分布中抽样

当两个连续事件的时间间隔 t 是随机的时候，仿真也就产生随机性. 本节将给出从一个概率分布 $f(t)$ 中产生连续的随机样本 $(t = t_1, t_2, \cdots)$ 的 3 种方法.

(1) 逆方法； (2) 卷积法； (3) 接受-舍弃法.

逆方法尤其适合于易于解析的概率密度函数, 如指数分布和均匀分布. 剩下的两种方法适合于求解一些较为复杂的情况, 例如正态分布和泊松分布. 所有这 3 种方法从根本上都用了独立同分布的均匀 (0,1) 随机数.

逆方法 假定现在需要从 (连续的或者离散的) 概率密度函数 $f(x)$ 中得到想要的随机样本 x. 逆方法首先对每一个定义的 y, 确定累积密度函数的闭的表达式 $F(x) = \mathrm{P}\{y \leqslant x\}$, 其中 $0 \leqslant F(x) \leqslant 1$. 已知 R 是一个从 (0,1) 均匀分布中得到的随机数, 并且假定 F^{-1} 表示 F 的逆, 则逆方法的步骤如下.

第 1 步 产生 (0,1) 上的随机数 R.

第 2 步 计算想要的样本, $x = F^{-1}(R)$.

图 16.5 直观地说明了产生一个连续的和一个离散的随机分布的方法. 均匀 (0,1) 随机数 R_1 从竖直方向上的 $F(x)$ 轴对应到水平轴上所需要的样本值 x_1.

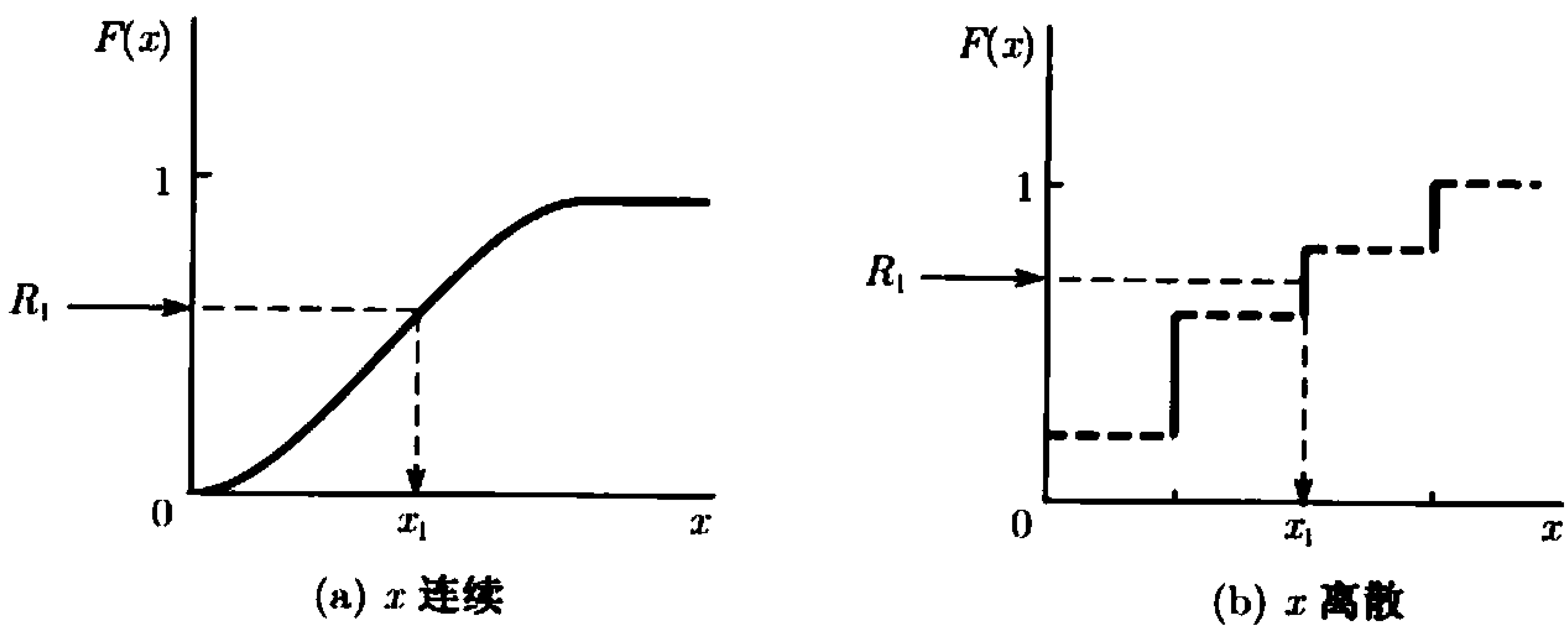


图 16.5 通过逆方法得到一个概率分布中的样本

上面方法的合理性在于, 随机变量 $z = F(x)$ 均匀地分布在区间 $0 \leqslant z \leqslant 1$ 上. 下面的定理将给出其证明.

定理 16.3-1 *给定随机变量 x 的累积密度函数 $F(x)$, $-\infty \leqslant x \leqslant \infty$, 随机变量 $z = F(x)$, $0 \leqslant z \leqslant 1$, 有下面的均匀 0-1 密度函数:*

$$f(z) = 1, \quad 0 \leqslant z \leqslant 1$$

证明 随机变量服从均匀分布当且仅当,

$$\mathrm{P}\{z \leqslant Z\} = Z, \quad 0 \leqslant Z \leqslant 1$$

此结果可应用到 $F(x)$, 由于:

$$\mathrm{P}\{z \leqslant Z\} = \mathrm{P}\{F(x) \leqslant Z\} = \mathrm{P}\{x \leqslant F^{-1}(Z)\} = F[F^{-1}(Z)] = Z$$

另外, 由于 $0 \leqslant \mathrm{P}\{z \leqslant Z\} \leqslant 1$, 所以 $0 \leqslant Z \leqslant 1$.

例 16.3-2 (指数分布)

指数概率密度函数

$$f(t) = \lambda \mathrm{e}^{-\lambda t}, t > 0$$

表示了顾客到达设备的到达时间间隔 t, 其期望是 $\frac{1}{\lambda}$. 根据 $f(t)$ 可以确定一个样本 t.

累积密度函数是:

$$F(t) = \int_0^t \lambda e^{-\lambda x} dx = 1 - e^{-\lambda t}, t > 0$$

令 $R = F(t)$, 可以解出 t, 得到

$$t = -\frac{1}{\lambda}\ln(1-R)$$

因为 $(1-R)$ 是 R 的补, 所以可以用 $\ln(R)$ 代替 $\ln(1-R)$.

考虑到仿真, 结果意味着顾客到达的时间间隔是 t 个时间单位. 例如, 对于每小时 $\lambda = 4$ 个客户和 $R = 0.9$, 计算直到下一个客户到达的时间间隔为:

$$t_1 = -\frac{1}{4}\ln(1-0.9) = 0.577 \text{ 小时} = 34.5 \text{ 分钟}$$

用于产生连续样本的 R 的值一定要随机地从 $(0,1)$ 均匀分布中抽取. 在接下来的 16.4 节中, 我们将详细介绍这些 $(0,1)$ 随机数在仿真过程中是如何产生的.

习题 16.3B

***1.** 在例题 16.3-2 中, 假定第一个客户在 0 时刻到达, 使用表 16.1 第一列的前 3 个随机数来产生接下来 3 个客户的到达时间, 并且在时间轴上画出相应的事件.

***2.** *均匀分布*. 假定加工一种机器的零件所需要的时间可以用下面的均匀分布来描述:

$$f(t) = \frac{1}{b-a}, \quad a \leqslant t \leqslant b$$

写出在给定随机数 R 的样本 t 的表达式.

3. 工件随机地到达有一台机器的车间, 到达时间服从期望是 2 小时的指数分布, 加工一个工件所需要的时间服从 1.1~2 小时的均匀分布. 假定第一个工件到达的时间为 0, 使用表 16.1 第一列中的 $(0,1)$ 随机数来确定前 5 个工件的到达和离开时间.

4. 对于某种客机的一个较为昂贵的零件的需求量, 每月为 0,1,2, 3 个单位的概率分别是 0.2, 0.3, 0.4, 0.1. 航空维修站开始的时候有这种零件 5 个单位的储备, 并且一旦当储备量低于 2 个单位则维修站将立刻增加储备到 5 个单位.

***(a)** 设计需求量抽样的方法.

(b) 直到发生第一次增加储备, 期间经历了多少个月? R 的连续值使用表 16.1 第一列中的数.

5. 在一个仿真过程中, 将检查电视的零件是否存在缺陷. 有 80% 的可能零件通过检查, 在这种情况下电视被送去包装. 否则, 将对电视进行维修. 我们用下面两种方法之一来象征性地描述这种情况.

去维修 /0.2, 包装 /0.8

去包装 /0.8, 维修 /0.2

这两种表示方法出现的机会是均等的. 但是, 当一个 (0, 1) 随机数序列应用到这两种表示方法时, 就会得到不同的决定 (维修或者包装). 请解释这是为什么.

6. 某人重复地投掷一枚均匀硬币直到正面朝上, 相应的盈利是 2^n, 其中 n 表示直到正面朝上一共投掷的次数.
   (a) 设计这个游戏的抽样方法.
   (b) 使用表 16.1 第一列中的随机数, 确定在第二次正面朝上时的累积盈利.
7. *三角分布*. 在仿真中, 由于缺乏数据, 可能使它无法确定与仿真活动相关的概率分布. 在大多数这种情况下, 很容易通过估算其最小值、平均值和最大值来描述所需的变量. 这 3 个值足以定义一个三角分布, 这就可以用作一个真实分布的粗略估计.
   (a) 根据下面的三角分布建立描述样本的方程, 相应的参数是 a, b, c:

$$f(x)=\begin{cases}\frac{2(x-a)}{(b-a)(c-a)}, & a\leqslant x\leqslant b\\ \frac{2(c-x)}{(c-b)(c-a)}, & b\leqslant x\leqslant c\end{cases}$$

   (b) 利用表 16.1 第一列的前 3 个随机数, 根据给定参数为 (1, 3, 7) 的三角分布来产生 3 个样本.
8. 下面考虑一个概率分布, 在一个矩形的两侧各有一个直角三角形, 其中左侧三角形、矩形、右侧三角形所在的区间分别是 $[a, b]$, $[b, c]$, $[c, d]$, $a < b < c < d$. 两个三角形与矩形有相同的高度.
   (a) 建立一个抽样程序.
   (b) 利用表 16.1 第一列的前 5 个随机数, 根据给定参数为 $(a, b, c, d) = (1, 2, 4, 6)$ 来产生 5 个样本.

*9. *几何分布*. 说明如何根据下面的几何分布来得到一个随机样本:

$$f(x)=p(1-p)^x, x=0,1,2,\cdots$$

参数 x 表示直到成功之前发生的 (伯努利试验) 失败的次数, p 表示成功的概率, $0 < p < 1$. 利用表 16.1 第一列的前 5 个随机数, 根据给定参数 $p = 0.6$ 来产生 5 个样本.

10. *韦布尔分布*.根据下面概率密度函数给出的韦布尔分布, 如何确定一个随机样本:

$$f(x)=\alpha\beta^{-\alpha}x^{\alpha-1}\mathrm{e}^{-(x/\beta)^{\alpha}}, x>0$$

其中 $\alpha > 0$ 为形状参数, $\beta > 0$ 为尺度参数.

卷积法　卷积法的基本思想是, 用其他一些易得样本的随机变量的统计和来表示所需要的样本. 这类分布中典型的有埃尔朗分布和泊松分布, 它们都是从指数分布抽样中得到的.

例 16.3-3 (埃尔朗分布)

m-埃尔朗随机变量定义为 m 个独立同分布的指数随机变量的统计和 (卷积). 用 y 表示一个 m-埃尔朗随机变量, 那么

$$y=y_1+y_2+\cdots+y_m$$

其中 $y_i, i=1,2,\cdots,m$, 是独立同分布的指数随机变量, 它们的概率密度函数定义为

$$f(y_i)=\lambda e^{-\lambda y_i},\quad y_i>0,\ i=1,2,\cdots,m$$

根据例 16.3-2, 第 i 个指数分布的一个样本可以表示为

$$y_i=-\frac{1}{\lambda}\ln\ R_i, i=1,2,\cdots,m$$

因此, 一个 m-埃尔朗样本可以表示为

$$y=-\frac{1}{\lambda}(\ln\ R_1+\ln\ R_2+\cdots+\ln\ R_m)=-\frac{1}{\lambda}\ln\left(\prod_{i=1}^{m}R_i\right)$$

为了详细说明上面式子的用法, 假定 $m=3,\lambda=4$ 个事件/小时, 利用表 16.1 第一列的前 3 个随机数, 可以得到 $R_1R_2R_3=(0.058\ 9)(0.673\ 3)(0.479\ 9)=0.019\ 0$, 即

$$y=-\frac{1}{4}\ln\ 0.019=0.991\text{小时}$$

例 16.3-4 (泊松分布)

12.3.1 节已经证明了: 如果连续事件发生的时间间隔服从指数分布, 那么每单位时间发生的事件数目服从的分布就是泊松分布, 反之亦然. 我们利用这种关系来对泊松分布进行抽样.

假定泊松分布的期望值是每单位时间 λ 个事件, 那么事件发生的时间间隔服从期望为 $\frac{1}{\lambda}$ 时间单位的指数分布. 这就意味着一个泊松样本 n 将在 t 时间单位内发生, 当且仅当,

$$\text{事件 } n \text{ 发生的时刻} \leqslant t < \text{事件 } n+1 \text{ 发生的时刻}$$

这种情形等同于

$$\begin{aligned} t_1+t_2+\cdots+t_n\leqslant t<t_1+t_2+\cdots+t_{n+1}, &\quad n>0\\ 0\leqslant t<t_1, &\quad n=0\end{aligned}$$

其中 $t_i, i=1,2,\cdots,n+1$, 来自于期望为 $\frac{1}{\lambda}$ 的指数分布的样本. 根据例 16.3-3 的结果, 可以得到

$$\begin{aligned} -\frac{1}{\lambda}\ln\left(\prod_{i=1}^{n}R_i\right)\leqslant t<-\frac{1}{\lambda}\ln\left(\prod_{i=1}^{n+1}R_i\right), &\quad n>0\\ 0\leqslant t<-\frac{1}{\lambda}\ln\ R_1, &\quad n=0\end{aligned}$$

简化为

$$\prod_{i=1}^{n} R_i \geqslant e^{-\lambda t} > \prod_{i=1}^{n+1} R_i,\ n>0$$

$$1 \geqslant e^{-\lambda t} > R_1, \qquad n=0$$

为了详细地描述这个抽样程序的执行, 假定 $\lambda = 4$ 个事件/小时, 并且我们希望得到时间段 $t = 0.5$ 小时内的一个样本. 可以得到 $e^{-\lambda t} = 0.135\,3$. 利用表 16.1 第一列中的随机数, 我们注意到 $R_1 = 0.058\,9$ 小于 $e^{-\lambda t} = 0.135\,3$. 因此相应的样本是 $n = 0$.

例 16.3-5 (正态分布)

中心极限定理 (见 14.4.4 节) 说明了, n 个独立同分布随机变量的和 (卷积) 当 n 足够大的时候近似于正态的. 我们将利用这个定理来产生正态分布的样本, 其中期望是 μ, 标准差是 σ.

定义

$$x = R_1 + R_2 + \cdots + R_n$$

根据中心极限定理可知, 这个随机变量近似于一个正态的. 给定服从 (0,1) 上均匀分布的随机数 R, 期望为 $\frac{1}{2}$, 方差为 $\frac{1}{12}$, 那么 x 的期望是 $\frac{n}{2}$, 方差为 $\frac{n}{12}$. 这样的话, 期望为 μ, 标准差为 σ 的正态分布 $N(\mu,\sigma)$ 的一个随机样本 y 可以根据 x 计算得到

$$y = \mu + \sigma\left(\frac{x-\frac{n}{2}}{\sqrt{\frac{n}{12}}}\right)$$

实际上, 为简便起见取 $n = 12$, 那么上式可以简化为

$$y = \mu + \sigma(x-6)$$

为了详细说明这个方法的应用, 假定我们希望得到分布 $N(10,2)$(期望为 $\mu = 10$, 标准差为 $\sigma = 2$) 的一个样本. 使用表 16.1 第 1 列和第 2 列中前 12 个随机数, 可以得到 $x = 6.109\,4$. 那么, $y = 10 + 2(6.109\,4 - 6) = 10.218\,8$.

这种方法的缺点就是, 对于每一个正态分布样本都需要产生 12 个随机数, 这导致计算的效率不高. 应用下面的转化以提高计算的效率:

$$x = \cos(2\pi R_2)\sqrt{-2\ln\ R_1}$$

Box and Muller(1958) 证明了这个 x 是一个标准正态分布 $N(0,1)$. 那么, $y = \mu+\sigma x$ 将产生一个 $N(\mu,\sigma)$ 的样本. 由于这个过程只需要两个 (0,1) 随机数, 因此它是比较有效的. 而实际上, 这种方法比上面说的更加有效, 因为 Box 和 Muller 已经证明了如果用 $\sin(2\pi R_2)$ 代替 $\cos(2\pi R_2)$, 则式子将可以给出另一个 $N(0,1)$ 样本.

下面来看一个 Box-Muller 方法应用的实际例子. 对于正态分布 $N(10,2)$, 使用表 16.1 第 1 列前两个随机数, 那么得到的 $N(0,1)$ 样本是:

$$x_1 = \cos(2\pi \times 0.673\ 3)\sqrt{-2\ln\ 0.058\ 9} \approx -1.103$$

$$x_2 = \sin(2\pi \times 0.673\ 3)\sqrt{-2\ln\ 0.058\ 9} \approx -2.109$$

那么相应的 $N(10,2)$ 样本是:

$$y_1 = 10 + 2(-1.103) = 7.794$$

$$y_2 = 10 + 2(-2.109) = 5.782$$

习题 16.3C[①]

*1. 在例 16.3-3 中, 给定 $m = 4$ 和 $\lambda = 4$ 个事件/小时, 计算一个埃尔朗样本.

2. 在例 16.3-4 中, 给定一个泊松分布的期望是 5 个事件/小时, 在 2 小时时间间隔内产生 3 个泊松样本.

3. 在例 16.3-5 中, 利用卷积法和 Box-Muller 方法产生两个 $N(8,1)$ 样本.

4. 工件到达 Metalco 加工厂服从泊松分布, 期望是每天 6 个工件. 工件达到之后, 按照一个严格的轮流顺序分配到 5 个加工中心. 确定到达第一个加工中心的工件的时间间隔的一个样本.

5. Springdale 高中的 1994 级学生的美国大学入学考试分数服从正态分布, 其中期望是 27 分, 标准差是 3 分, 假定我们需要得到这个班一个 6 人的随机抽样. 利用 Box-Muller 方法确定这个抽样的期望和标准差.

6. 心理学教授 Yataha 在进行学习实验中训练小鼠绕迷宫, 该迷宫是一个方形的. 小鼠在其中的四个角落之一进入迷宫, 而且必须找到一条穿过迷宫的路, 并且从进入的口离开迷宫. 所设计的迷宫使得小鼠必须通过其余 3 个角点恰好一次再离开. 连接迷宫中 4 个角点的多条路径都是严格的顺时针方向, Yataha 教授估计小鼠从一个角点到另一个角点的时间服从 10 秒到 20 秒的均匀分布, 这个时间依赖于小鼠走的路线. 为小鼠在迷宫中花费的时间设计一个抽样方法.

7. 假定在第 6 题中, 当一只小鼠从迷宫中出来时, 另外一只小鼠立刻进入迷宫. 设计一个 5 分钟内离开迷宫的小鼠数目的抽样方法.

8. *负二项分布*.说明如何根据下面的负二项分布来确定一个随机抽样:

$$f(x) = C_x^{r+x-1}p^r(1-p)^x,\quad x = 0,1,2,\cdots$$

其中 x 表示在独立伯努利试验序列中直到第 r 次成功试验之前试验失败的次数, p 是试验成功的概率, $0 < p < 1$. (*提示*: 负二项分布是 r 个独立几何抽样的卷积, 见 16.3B 第 9 题.)

接受-舍弃法 接受-舍弃法针对的是较为复杂且不能够用前面所给出的方法解决的一些概率密度函数. 这种方法的一般思想是, 用一个较为容易解析处理的概率密

① 对于本节的所有习题, 使用的随机数都是从表 16.1 中第 1 列开始的.

度函数 $h(x)$ 来替代这个复杂的概率密度函数 $f(x)$, 那么从 $h(x)$ 得出的抽样就可以用来作为原始概率密度 $f(x)$ 的样本.

定义**控制函数** $g(x)$ 使得它可以在整个区间上控制 $f(x)$, 即

$$g(x) \geqslant f(x), \quad -\infty < x < \infty$$

然后, 定义替代概率密度函数 $h(x)$, 通过单位化 $g(x)$ 得到:

$$h(x) = \frac{g(x)}{\int_{-\infty}^{\infty} g(y)\mathrm{d}y}, \quad -\infty < x < \infty$$

下面给出接受-舍弃法的基本步骤.

第 1 步 利用逆方法或者卷积法从 $h(x)$ 中得到一个样本 $x = x_1$.

第 2 步 得到一个 $(0,1)$ 随机数 R.

第 3 步 如果 $R \leqslant \frac{f(x_1)}{g(x_1)}$, 那么接受 x_1 作为 $f(x)$ 的一个样本. 否则, 舍弃 x_1 并且返回到第 1 步.

这个方法的合理性是基于下面的等式:

$$\mathrm{P}\{x \leqslant a | x = x_1\text{被接受}, -\infty < x_1 < \infty\} = \int_{-\infty}^{a} f(y)\mathrm{d}y, \quad -\infty < a < \infty$$

这个概率的命题说明了, 现实中满足第 3 步条件的样本 $x = x_1$ 是原始概率密度函数 $f(x)$ 的一个样本.

可以通过减小第 3 步中的舍弃概率来提高上面给出的方法的效率, 这个概率依赖于控制函数 $g(x)$ 的具体选择, 并且随着选取的 $g(x)$ 更贴近 $f(x)$, 这个概率会减小.

例 16.3-6 (β 分布)

将接受-舍弃法应用到下面的 β 分布:

$$f(x) = 6x(1-x), \quad 0 \leqslant x \leqslant 1$$

图 16.6 描述了 $f(x)$ 和控制函数 $g(x)$.

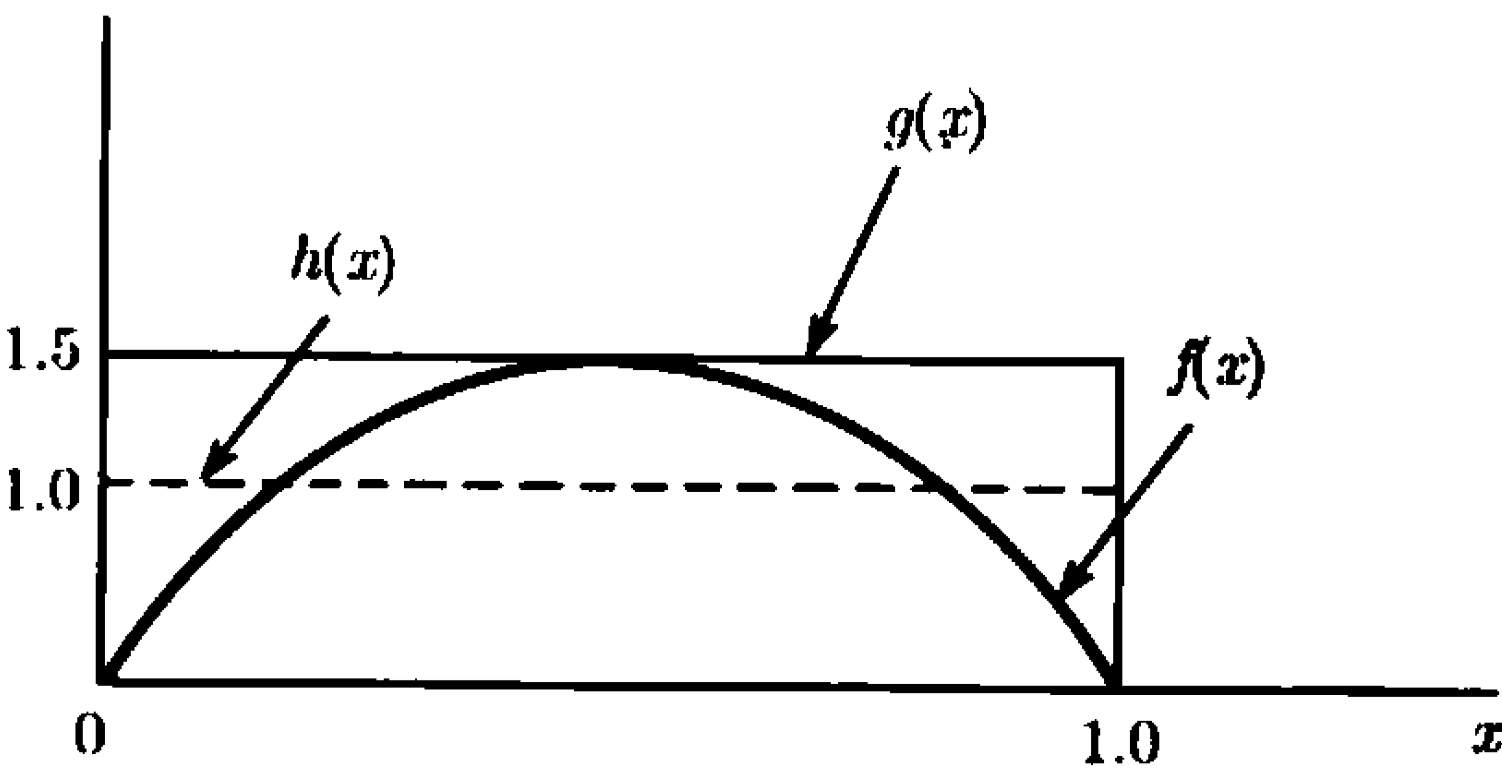


图 16.6 β 分布 $f(x)$ 的控制函数 $g(x)$

控制函数 $g(x)$ 的高度等于函数 $f(x)$ 的最大值, 并且当 $x = 0.5$ 时取到最大值. 因此矩形的高度等于 $f(0.5) = 1.5$. 这意味着

$$g(x) = 1.5,\ 0 \leqslant x \leqslant 1$$

图 16.6 中也给出了替代概率密度函数 $h(x)$ 的计算结果:

$$h(x) = \frac{g(x)}{g(x)\text{下方的区域}} = \frac{1.5}{1 \times 1.5} = 1,\ 0 \leqslant x \leqslant 1$$

使用表 16.1 的 $(0,1)$ 随机数列来详细说明这个过程.

第 1 步 $R = 0.058\ 9$ 给出了 $h(x)$ 的一个样本 $x = 0.058\ 9$.

第 2 步 $R = 0.673\ 3$.

第 3 步 因为 $\frac{f(0.058\ 9)}{g(0.058\ 9)} = \frac{0.332\ 6}{1.5} = 0.221\ 7$ 小于 $R = 0.673\ 3$, 所以接受样本 $x_1 = 0.058\ 9$.

为了得到第二个样本, 我们如下继续.

第 1 步 $R = 0.479\ 9$ 给出了 $h(x)$ 的另一个样本 $x = 0.479\ 9$.

第 2 步 $R = 0.948\ 6$.

第 3 步 因为 $\frac{f(0.479\ 9)}{g(0.479\ 9)} = 0.998\ 4$ 大于 $R = 0.948\ 6$, 所以拒绝 $x = 0.479\ 9$ 作为一个合理的 β 样本. 这就意味着我们需要利用新的随机数来重复这些步骤直到第 3 步的条件得到满足.

评注 接受-舍弃法可以通过选取能够更紧地包含 $f(x)$ 的一个控制函数 $g(x)$ 来提高方法的效率, 这样也就可以得到一个更加易于解析的替代函数 $h(x)$. 例如, 在图 16.6 中, 如果用阶梯金字塔函数 (参见 16.3D 第 2 题) 代替矩形控制函数 $g(x)$, 就可以使得接受-舍弃法更加有效. 阶梯数越大, $g(x)$ 逼近 $f(x)$ 就越好, 因此一个样本被接受的概率也就越高. 然而, 越紧的控制函数通常也会带来计算量的增大, 甚至会抵消掉被接受概率的增加量.

习题 16.3D

1. 在例 16.3-5 中, 继续程序的步骤直到得到一个可行的样本. 使用表 16.1 中的 $(0,1)$ 随机数, 并且与例题中使用的顺序相同.
2. 考虑例 16.3-6 中的 β 概率密度函数. 确定一个两阶梯金字塔控制函数 $g(x)$, 两个阶梯的高度都是 $\frac{1.5}{2} = 0.75$. 使用表 16.1 中的. 与例 16.3-6 中使用的相同的 $(0,1)$ 随机数序列, 根据这个新的控制函数, 获得一个 β 抽样. 一般的结论是, 较紧的控制函数可以提高抽样被接受的概率. 同时还要注意的是, 对应于新函数的计算量也会增大.
3. 对于下面的函数使用接受-舍弃法, 确定相应的函数 $g(x)$ 和 $h(x)$:

$$f(x) = \frac{\sin(x) + \cos(x)}{2},\quad 0 \leqslant x \leqslant \frac{\pi}{2}$$

使用表 16.1 第 1 列的 $(0,1)$ 随机数来产生 $f(x)$ 的两个样本. [提示: 为简便起见, 对于 $f(x)$ 已定义的区域, 使用矩形函数 $g(x)$ 来覆盖.]

4. 在 HairKare 理发店, 顾客到达的时间间隔服从下面的分布:

$$f_1(t)=\frac{k_1}{t},\quad 12\leqslant t\leqslant 20$$

理发的时间服从下面的分布:

$$f_2(t)=\frac{k_2}{t^2},\quad 18\leqslant t\leqslant 22$$

其中, 常数 k_1 和 k_2 使得函数 $f_1(t)$ 和 $f_2(t)$ 成为概率密度函数. 利用接受–舍弃法 (以及表 16.1 中的随机数) 来确定第一个顾客离开 HairKare 的时间, 以及下一个顾客什么时间到达. 假定第一个顾客到达的时间是 $T=0$.

16.4 随机数的生成

在分布抽样过程中, 均匀 (0, 1) 随机数扮演了非常重要的角色. 真正的 (0, 1) 随机数只能通过电子装置来产生. 然而, 由于仿真试验是在计算机上实现的, 所以使用电子装置来产生随机数对于试验目的就显得太慢. 再者, 电子装置是通过机会定律 (law of chance) 来激活的, 因此它不能够随心所欲地复制同一个随机数序列, 这一点是非常重要的, 因为在仿真模型中的调试、验证和确认的步骤通常需要复制同一个随机数序列.

产生仿真试验中所使用的 (0, 1) 随机数的唯一合理的方法是基于算术运算的. 由于这样的数是提前产生的, 所以它们并不是真正的随机数. 因此, 称这种数为**伪随机数** (pseudo-random numbers) 是比较确切的.

最为常用的产生 (0, 1) 随机数的算术运算方法是**乘同余法** (multiplicative congruential method). 给定参数 u_0, b, c, m, 一个伪随机数序列 R_n 可以根据下面的式子得到:

$$u_n=(bu_{n-1}+c)\ \mathrm{mod}(m),\quad n=1,2,\cdots$$

$$R_n=\frac{u_n}{m},\quad n=1,2,\cdots$$

初始的值 u_0 通常成为随机数生成器的**种子** (seed).

Law and Kelton(1991) 改进了乘同余法, 提高了随机数生成器的性能.

例 16.4-1

使用参数 $b=9, c=5, m=12$, 使用乘同余法产生 3 个随机数. 种子是 $u_0=11$.

$$u_1=(9\times 11+5)\ \mathrm{mod}\ 12=8,\ R_1=\frac{8}{12}=0.666\ 7$$

$$u_2=(9\times 8+5)\ \mathrm{mod}\ 12=5,\ R_2=\frac{5}{12}=0.416\ 7$$

$$u_3=(9\times 5+5) \bmod 12=2,\quad R_3=\frac{2}{12}=0.166\,7$$

Excel 程序

文件 excelRN.xls 包含了使用乘同余法计算随机数的 Excel 模板. 图 16.7 给出了与例 16.3-1 中参数有关的随机数序列. 仔细观察会发现随机数循环的长度是 4, 即经过 4 个随机数后这个序列又是它自身. 据此可以看出, 参数 u_0, b, c, m 的选择直接决定了生成器产生随机数的 (统计) 质量, 以及随机数序列循环的长度. 因此, 不建议 "偶然地" 使用乘同余法的式子来得到随机数, 而应当使用一个可靠的并且已经测试过的生成器. 实际上所有的商用计算机程序都安装了可靠的随机数生成器.

Multiplicative Congruential Method	
Input data(B7<=1000)	
b =	9
c =	5
u0 =	11
m =	12
How many numbers?	10
Output results	
Generated random numbers:	
1	0.66667
2	0.41667
3	0.16667
4	0.91667
5	0.66667
6	0.41667
7	0.16667
8	0.91667
9	0.66667
10	0.41667

图 16.7 根据例 16.4-1 中数据得到的 Excel 随机数 (文件 excelRN.xls)

习题 16.4A

*1. 根据下面的参数使用模板 excelRN.xls 得到随机数序列, 并且与例 16.4-1 中的结果进行比较:

$$b=17,\quad c=111,\quad m=103,\quad \text{种子}=7$$

2. 找出你计算机上的一个随机数生成器, 并且使用它产生 500 个 (0, 1) 随机数. 用直方图检验得出的结果 (可以使用微软的直方图工具, 参见 14.5 节), 并确信得到的随机数服从 (0, 1) 均匀分布. 为了更确切地分析这些随机数, 可以使用下面的试验: χ^2 拟合优度检验 (参见 14.5 节)、独立性检验、相关性检验 (详见 Law and Kelton[1991]).

16.5 离散仿真的方法

本节将详细介绍在仿真模型中如何收集典型的统计信息, 并且将使用一个单排队模型加以说明. 16.5.1 节使用一个数值实例来详细说明单服务台排队仿真模型中的计算过程和方法. 由于仿真模型中的计算量相当大, 所以 16.5.2 节将给出如何使用 Excel 电子表格来完成单服务台排队模型的建模和计算.

16.5.1 单服务台模型的人工仿真

HairKare 理发店的顾客到达时间间隔服从期望为 15 分钟的指数分布. 理发店只有一名理发师, 他为顾客理发的时间服从 10~15 分钟的均匀分布. 顾客接受服务的顺序服从先到先服务的规则 (FIFO). 仿真的目标是计算出下面的几个量.

(1) 理发店的平均使用时间. (2) 等待顾客的平均数目.

(3) 顾客在队列中的平均等待时间.

仿真模型的逻辑性可以使用相应的到达事件和离开事件的行为加以描述.

到达事件

(1) 依次产生并存储下一个到达事件的发生时间 (= 当前仿真时间 + 到达间隔时间).

(2) 如果设备 (理发师) 是空闲的:

(a) 开始服务并且声明设备为忙的, 更新设备使用统计表.

(b) 依次产生并存储顾客离开事件的时间 (= 当前仿真时间 + 服务时间).

(3) 如果设备是忙的, 那么顾客进入等候队列, 并更新队列统计表.

离开事件

(1) 如果队列为空, 那么声明设备是空闲的, 并更新设备使用统计表.

(2) 如果队列不是空的:

(a) 从队列中选择一位顾客, 接受设备的服务, 并更新队列和设备使用统计表.

(b) 依次产生并存储顾客离开事件的时间 (= 当前仿真时间 + 服务时间).

根据这个问题提供的数据可以知道, 到达时间间隔服从期望为 15 分钟的指数分布, 服务时间服从 10~15 分钟的均匀分布. 用 p 和 q 分别表示到达时间间隔和服务时间的随机抽样, 那么, 根据 16.3.2 节中的解释可以得到

$$p=-15\ln R \text{ 分钟}, \quad 0\leqslant R\leqslant 1$$

$$q=10+5R \text{ 分钟}, \quad 0\leqslant R\leqslant 1$$

为了求解例题的需要, 对于 R 我们使用表 16.1 中的随机数, 从第一列开始. 同时我们用符号 T 表示仿真的时钟时间, 进一步假定: 第一个顾客到达的时间为 $T=0$, 并且服务设施初始的时候是空闲的.

由于仿真的计算量相当大, 所以我们只计算前 5 位顾客. 我们设计的例题可以涵盖仿真过程中的所有可能的情形. 本节随后还将介绍 excelSingleServer.xls 模板, 它可以让我们在不需要人工完成计算的条件下来执行仿真模型.

第 1 位顾客在 $T = 0$ 时刻到达 产生第 2 位顾客的到达时间:

$$T = 0 + p_1 = 0 + (-15\ln\ 0.058\ 9) = 42.48\ \text{分钟}$$

由于在 $T = 0$ 时刻设备是空闲的, 所以第 1 位顾客立刻开始接受服务, 因此这位顾客的离开时间为:

$$T = 0 + q_1 = 0 + (10 + 5 \times 0.673\ 3) = 13.37\ \text{分钟}$$

因此可以得到将来事件的顺序表:

时刻 T	事　件
13.37	第 1 位顾客离开
42.48	第 2 位顾客到达

第 1 位顾客在 $T = 13.37$ 离开 因为队列为空, 所以服务设施声明为空闲. 同时, 我们记录下设备从 $T = 0$ 到 $T = 13.37$ 是忙的. 更新将来事件的顺序表为

时刻 T	事　件
42.48	第 2 位顾客到达

第 2 位顾客在 $T = 42.48$ 到达 第 3 位顾客的到达时间是

$$T = 42.48 + (-15\ln\ 0.479\ 9) = 53.49\ \text{分钟}$$

由于此时设备是空闲的, 所以第 2 位顾客立刻开始接受服务, 并且声明设备是忙的. 离开时间是:

$$T = 42.48 + (10 + 5 \times 0.948\ 6) = 57.22\ \text{分钟}$$

更新将来事件的顺序表为

时刻 T	事　件
53.49	第 3 位顾客到达
57.22	第 2 位顾客离开

第 3 位顾客在 $T = 53.49$ 到达 第 4 位顾客的到达时间是

$$T = 53.49 + (-15\ln\ 0.613\ 9) = 60.81\ \text{分钟}$$

因为设备当前是忙的 (直到 $T = 57.22$), 第 3 位顾客在 $T = 53.49$ 加入队列. 更新将来事件的顺序表为:

时刻 T	事件
57.22	第 2 位顾客离开
60.81	第 4 位顾客到达

第 2 位顾客在 $T = 57.22$ 离开 第 3 位顾客走出队列, 开始接受服务, 他的等待时间为:

$$W_3 = 57.22 - 53.49 = 3.73 \text{ 分钟}$$

离开时间为:

$$T = 57.22 + (10 + 5 \times 0.593\,3) = 70.19 \text{ 分钟}$$

更新将来事件的顺序表为:

时刻 T	事件
60.81	第 4 位顾客到达
70.19	第 3 位顾客离开

第 4 位顾客在 $T = 60.81$ 到达 第 5 位顾客的到达时间是:

$$T = 60.81 + (-15 \ln\ 0.934\,1) = 61.83 \text{ 分钟}$$

因为设备一直到 $T = 70.19$ 都是忙的, 所以第 4 位顾客进入等候队列. 更新将来事件的顺序表为:

时刻 T	事件
61.83	第 5 位顾客到达
70.19	第 3 位顾客离开

第 5 位顾客在 $T = 61.83$ 到达 因为我们只考虑前 5 位顾客到达, 因此不会产生第 6 位顾客的到达事件. 设备在 $T = 61.83$ 仍旧很忙, 所以第 5 位顾客需要进入队列等候. 更新将来事件的顺序表为:

时刻 T	事件
70.19	第 3 位顾客离开

第 3 位顾客在 $T = 70.19$ 离开 第 4 位顾客走出队列开始接受服务, 他的等待时间为:

$$W_4 = 70.19 - 60.81 = 9.38 \text{ 分钟}$$

离开时间为:

$$T = 70.19 + (10 + 5 \times 0.178\,2) = 81.08 \text{ 分钟}$$

更新将来事件的顺序表为:

时刻 T	事件
81.08	第 4 位顾客离开

第 4 位顾客在 $T = 81.08$ 离开 第 5 位顾客走出队列开始接受服务，他的等待时间为：

$$W_5 = 81.08 - 61.83 = 19.25 \text{ 分钟}$$

离开时间为：

$$T = 81.08 + (10 + 5 \times 0.347\,3) = 92.82 \text{ 分钟}$$

更新将来事件的顺序表为：

时刻 T	事　件
92.82	第 5 位顾客离开

第 5 位顾客在 $T = 92.82$ 离开 此时系统中 (队列中和设备上) 没有顾客了，仿真结束.

图 16.8 直观地给出了队列长度以及设备使用作为仿真时间的函数的变化情况.

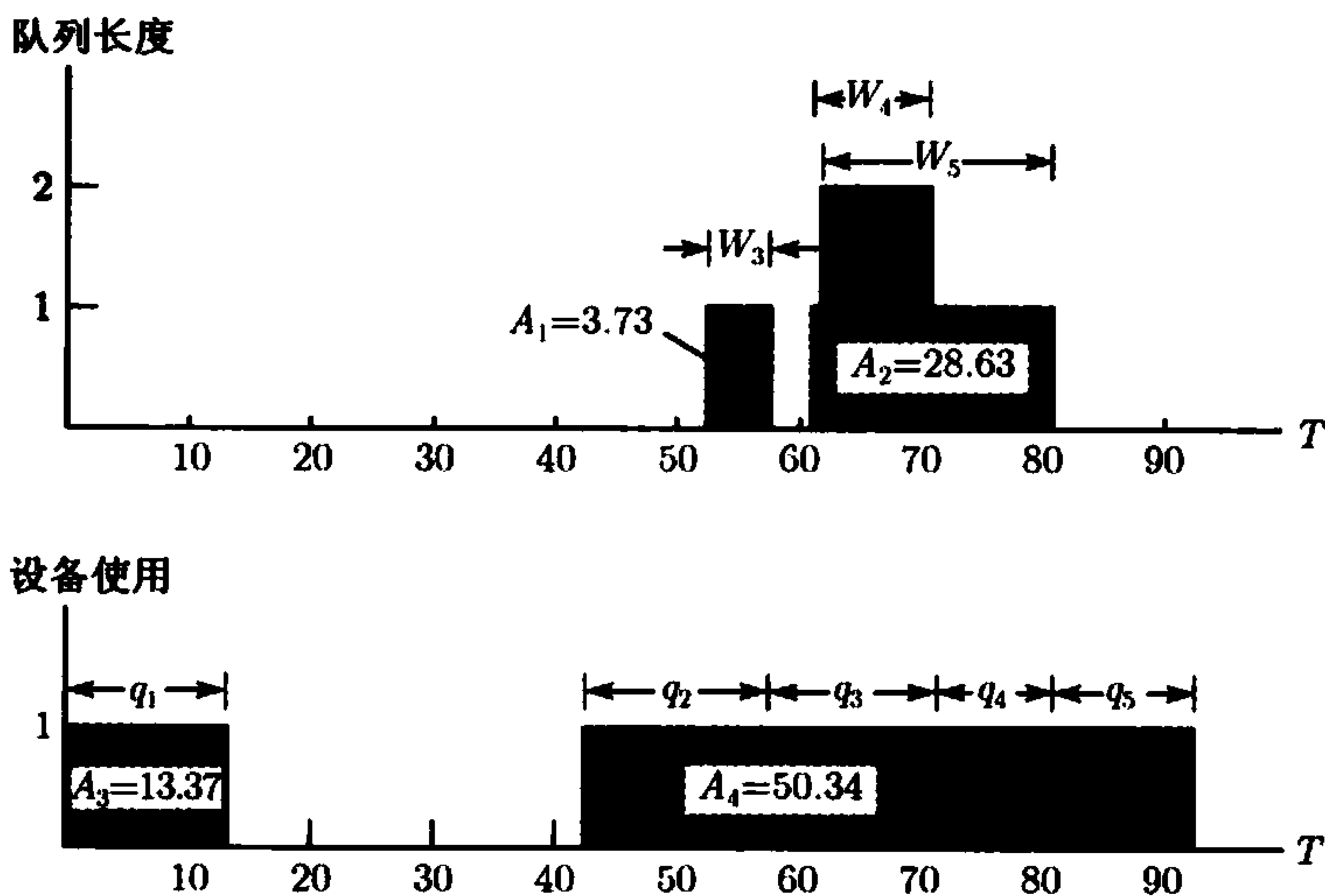


图 16.8 队列长度以及设备使用作为仿真时间的函数的变化情况

队列长度和设备使用是基于时间的变量，因为它们的变化是时间的函数. 因此，可以计算出它们的平均值：

$$\begin{pmatrix}\text{一个基于时间的}\\\text{变量的平均值}\end{pmatrix} = \frac{\text{曲线以下的面积}}{\text{仿真时间}}$$

根据图 16.8 给定的数据运用上面的式子，可以得到

$$\begin{pmatrix}\text{队列的}\\\text{平均长度}\end{pmatrix} = \frac{A_1 + A_2}{92.82} = \frac{32.36}{92.82} = 0.349 \text{ 顾客}$$

$$\begin{pmatrix}\text{设备平均}\\\text{使用率}\end{pmatrix} = \frac{A_3 + A_4}{92.82} = \frac{63.71}{92.82} = 0.686 \text{ 理发师}$$

顾客在队列中的平均等待时间是一个基于观测的变量, 可以根据下面的式子确定:

$$\begin{pmatrix}\text{一个基于观测的}\\ \text{变量的平均值}\end{pmatrix}=\frac{\text{观测值的总和}}{\text{观测次数}}$$

仔细观察图 16.8 可以发现, 队列长度曲线下方的区域面积, 恰好等于加入到队列中的 3 位顾客的等待时间总和, 也就是,

$$W_1+W_2+W_3+W_4+W_5=0+0+3.73+9.38+19.25=32.36 \text{ 分钟}$$

队列中所有顾客的平均等待时间是

$$\overline{W_q}=\frac{32.36}{5}=6.47 \text{ 分钟}$$

习题 16.5A

1. 现在假定 16.5.1 节中理发店有两名理发师, 顾客按照先到先服务 (FCFS) 的规则. 进一步假定为一名顾客理发的时间服从 15~30 分钟的均匀分布. 顾客到达的时间间隔服从期望是 10 分钟的指数分布. 人工仿真系统运行 75 个时间单位. 根据仿真的结果, 确定一个顾客在队列中的平均等待时间, 顾客的平均等待数目, 理发师的平均服务时间. 使用表 16.1 中的随机数.
2. 将下面的变量分类, 或者是基于观测的或者是基于时间的:
   *(a) 一个电子元件的寿命.
   *(b) 一种物品的库存水平.
   (c) 一种库存物品的订货量.
   (d) 一批物品中次品的数目.
   (e) 评卷需要的时间.
   (f) 一家汽车租赁公司的停车场中汽车的数目.

*3. 下面的表格表示, 在队列中等待的顾客数目是仿真时间的函数.

仿真时间 T(小时)	等待的顾客的数目	仿真时间 T(小时)	等待的顾客的数目
$0\leqslant T\leqslant 3$	0	$10<T\leqslant 12$	2
$3<T\leqslant 4$	1	$12<T\leqslant 18$	3
$4<T\leqslant 6$	2	$18<T\leqslant 20$	2
$6<T\leqslant 7$	1	$20<T\leqslant 25$	1
$7<T\leqslant 10$	0		

求出下面的各个量.

(a) 队列的平均长度.
(b) 必须在队列中等待的顾客的平均等待时间.

4. 假定 16.5.1 节中的理发店有 3 名理发师, 下表给出了理发师在各个时间段的工作情况:

仿真时间 T(小时)	忙碌的理发师的数目	仿真时间 T(小时)	忙碌的理发师的数目
$0<T\leqslant 10$	0	$60<T\leqslant 70$	2
$10<T\leqslant 20$	1	$70<T\leqslant 80$	3
$20<T\leqslant 30$	2	$80<T\leqslant 90$	1
$30<T\leqslant 40$	0	$90<T\leqslant 100$	0
$40<T\leqslant 60$	1		

确定下面各个量的值.

(a) 忙碌的理发师的平均数目. (b) 理发师的平均忙碌时间.

(c) 理发师的平均空闲时间.

16.5.2 单服务台模型的电子表格仿真

根据 16.5.1 节, 仿真计算的过程通常非常乏味和繁琐, 因此, 有必要利用计算机执行仿真模型. 本节将介绍一个基于电子表格的模型来完成单服务台模型的仿真, 目的是更加深刻地理解 16.5.1 节中仿真的思想. 当然, 单服务台模型是一种非常简单的情形, 正是由于这个原因它更易于在电子表格环境下执行. 其他的一些较为复杂的情形则需要更多的考虑建模的方法, 可以利用一些现有的仿真软件包 (见 16.7 节).

16.5.1 节介绍了单服务台仿真模型中有两个基本要素:

(1) 模型中的事件顺序表.

(2) 跟踪设备使用情况和队列长度的变化的图表.

这两个要素在基于电子表格 (实际上是任何的基于计算机) 的仿真模型中也是非常重要的, 唯一的不同就是使得执行的规则可以让计算机来识别. 正如 16.5.1 节所说, 顾客接受服务是按照先到先服务的规则.

图 16.9 显示出了文件 excelSingleServer.xls 的结果. 输入的数据可以用下面 4 种分布之一表示到达间隔时间和服务时间: 常数分布、指数分布、均匀分布和三角分布. 当只提供 3 个估计数 a, b, c 分别代表一个分布中的间隔时间或者服务时间的最小值、最有可能值和最大值的时候, 三角分布可以给出这个分布的一个粗略的初始估计. 然后启动仿真程序所需要的其他信息就只有仿真执行的长度, 在本例中这个长度就是仿真模型所生成的到达顾客的数目.

电子表格的计算为每一位顾客的到达保留一行. 每位顾客到达的时间间隔和服务时间由输入数据生成. 第一位顾客到达发生在 $T=0$, 因为设备初始的时候是空闲的, 所以顾客立刻接受服务. 因此,

$$\begin{pmatrix}\text{第 1 位顾客的}\\\text{离开时间}\end{pmatrix}=\begin{pmatrix}\text{第 1 位顾客的}\\\text{到达时间}\end{pmatrix}+\begin{pmatrix}\text{第 1 位顾客的}\\\text{服务时间}\end{pmatrix}$$
$$=0+12.83=12.83$$
$$\begin{pmatrix}\text{第 2 位顾客的}\\\text{到达时间}\end{pmatrix}=\begin{pmatrix}\text{第 1 位顾客的}\\\text{到达时间}\end{pmatrix}+\begin{pmatrix}\text{与第 1 位顾客}\\\text{的到达间隔时间}\end{pmatrix}$$
$$=0+3.37=3.37$$

可以利用下面的式子确定任一顾客 i 的离开时间:

$$\begin{pmatrix}\text{顾客 } i \text{ 的}\\\text{离开时间}\end{pmatrix}=\max\left\{\begin{pmatrix}\text{顾客 } i \text{ 的}\\\text{到达时间}\end{pmatrix},\begin{pmatrix}\text{顾客 } i-1 \text{ 的}\\\text{离开时间}\end{pmatrix}\right\}+\begin{pmatrix}\text{顾客 } i \text{ 的}\\\text{服务时间}\end{pmatrix}$$

从上面的式子可以知道, 只有当设备空闲时, 才能服务顾客. 在图 16.9 上应用上面的式子可以得到:

$$\text{第 3 位顾客的离开时间} = \max\{9.09, 27.55\} + 12.21 = 39.76$$

Simulation of a Single-Server Queueing Model

Nbr of arrivals =		20	<<Maximum 500				
Enter x in column A to select interarrival pdf:							
	Constant =						
x	Exponential:	λ =	0.067				
	Uniform:	a =		b =			
	Triangular:	a =		b =		c =	
Enter x in column A to select service time pdf:							
	Constant =						
	Exponential:	μ =					
x	Uniform:	a =	10	b =	15		
	Triangular:	a =		b =		c =	
Output Summary							
Av. facility utilization =			0.98	Press F9 to trigger a new simulation run.			
Percent idleness (%) =			1.95				
Av. queue length, Lq =			1.57				
Av. nbr in system, Ls =			2.55				
Av. queue time, Wq =			21.24				
Av. system time, Ws =			34.47				
[illegible] =			[illegible]				
[illegible]			[illegible]				
[illegible]			[illegible]				

Nbr	InterArvlT	ServiceT	ArvlT	DepartT	Wq	Ws
1	3.73	12.83	0.00	12.83	0.00	12.83
2	5.37	14.71	3.73	27.55	9.10	23.82
3	3.86	12.21	9.09	39.75	18.45	30.66
4	14.10	11.18	12.95	50.94	26.80	37.98
5	7.35	14.92	27.05	65.85	23.88	38.80
6	35.70	14.22	34.41	80.07	31.45	45.67
7	0.60	14.50	70.11	94.58	9.97	24.47
8	4.25	13.35	70.71	107.93	23.87	37.22
9	4.85	12.45	74.96	120.38	32.97	45.41
10	7.43	11.57	79.81	131.94	40.56	52.13
11	8.99	14.65	87.24	146.59	44.70	59.34
12	49.78	12.85	96.23	159.43	50.36	63.20
13	0.42	14.12	146.01	173.55	13.43	27.54
14	8.77	13.69	146.43	187.24	27.13	40.82
15	11.19	10.50	155.20	197.75	32.05	42.55
16	42.82	13.78	166.38	211.53	31.36	45.14
17	19.87	12.29	209.20	223.82	2.33	14.62
18	9.25	12.95	229.07	242.03	0.00	12.95
19	13.98	12.99	238.33	255.02	3.70	16.69
20	58.46	14.88	252.31	269.90	2.71	17.59

图 16.9 单服务台仿真模型的 Excel 输出结果 (文件 excelSingleServer.xls)

下面讨论如何收集模型的统计数据. 首先, 对于顾客 i, 他在队列中的等待时间 $W_q(i)$, 以及在整个系统中的等待时间 $W_s(i)$, 可以由下面的式子计算给出:

$$W_q(i) = \begin{pmatrix}\text{顾客 } i \text{ 的}\\ \text{离开时间}\end{pmatrix} - \begin{pmatrix}\text{顾客 } i \text{ 的}\\ \text{到达时间}\end{pmatrix} - \begin{pmatrix}\text{顾客 } i \text{ 的}\\ \text{服务时间}\end{pmatrix}$$

$$W_s(i) = \begin{pmatrix}\text{顾客 } i \text{ 的}\\ \text{离开时间}\end{pmatrix} - \begin{pmatrix}\text{顾客 } i \text{ 的}\\ \text{到达时间}\end{pmatrix}$$

然后, 跟踪设备的使用情况以及队列长度的变化对于计算模型中其他的统计信息很有必要 (参见 16.5.1 节的过程), 幸运的是我们已经根据 16.5.1 节中的观测和图 16.8 的解释简化了这种计算:

(1) 设备使用曲线下方的面积 = 所有顾客的总的服务时间

(2) 队列长度曲线下方的面积 = 所有顾客等待时间的总和

为了说明这一点, 图 16.9 的 Excel 输出计算了这两个总和:

$$\text{服务时间总和} = 264.65$$

$$W_q\text{的总和} = 424.8$$

$$W_s\text{的总和} = W_q\text{的和} + \text{服务时间总和} = 689.44 (= 264.65 + 424.8)$$

给定最后一位顾客 (第 20 位顾客) 的离开时间 $T = 269.90$, 可以得到:

$$\begin{pmatrix}\text{设备平均}\\ \text{使用率}\end{pmatrix} = \frac{264.65}{269.90} = 0.980\,5$$

$$\begin{pmatrix}\text{队列的}\\ \text{平均长度}\end{pmatrix} = \frac{424.80}{269.90} = 1.57$$

设备空闲时间比例是 $(1 - 0.98) \times 100 = 1.945\%$

其他统计信息可以直接得到, 也就是

$$\begin{pmatrix}\text{在队列中的}\\ \text{平均等待时间}\end{pmatrix} = \frac{W_q\text{的总和}}{\text{总的顾客数}} = \frac{424.80}{20} = 21.24$$

$$\begin{pmatrix}\text{在系统中的}\\ \text{平均等待时间}\end{pmatrix} = \frac{W_s\text{的总和}}{\text{总的顾客数}} = \frac{689.44}{20} = 34.47$$

为了仿真多服务台模型, 人们设计了另一个电子表格 (excelMultiServer.xls). 模板的设计思想与单服务台相同, 但是在计算离开时间时有所不同, 因此需要使用 VBA 宏.

习题 16.5B

1. 使用 16.5.1 节中的输入数据, 对于 10 位顾客, 运行 Excel 模拟器, 并且用图表显示出设备使用率和队列长度作为仿真时间的函数的变化. 证明曲线下方的面积分别等于服务时间的总和以及等待时间的总和.
2. 已知顾客到达率为每小时 $\lambda = 4$ 位顾客, 服务率为每小时 $\mu = 6$ 位顾客离开, 仿真 500 位顾客到达的 $M/M/1$ 模型. 仿真执行 5 次 (通过更新电子表格 —— 按 F9), 并且给模型中各个量确定一个 95% 的置信区间. 将得到的结果与 $M/M/1$ 模型的稳态理论值进行比较.
3. 在单个操作台上, 电视元件从传送带上每 15 分钟到达这个操作台接受检查. 具体操作台检查结果不详. 然而操作员预计检查一个元件的平均时间为 10 分钟, 在最坏情况下的检查时间也不会超过 13 分钟, 而对于某些元件的检查时间则可以低到 9 分钟.
   (a) 使用 Excel 模拟器, 仿真 200 个电视元件的检查过程.
   (b) 根据重复 5 次仿真过程的结果, 估计等待检查的元件的平均数目以及操作台的平均使用率.

16.6 收集统计观测数据的方法

仿真是一种统计实验, 其输出结果必须使用恰当的统计推断工具 (例如, 置信区间与假设检验) 加以分析. 为了完成这个任务, 仿真实验中的观测数据必须满足下面的 3 个条件:

(1) 观测是来自平稳 (相同) 的分布.

(2) 观测从正态总体中采样. (3) 观测是独立的.

从严格意义上讲, 仿真实验不可能完全满足上面的 3 个条件中的任何一个. 然而, 可以使用仿真观测数据收集的规则来保证这些条件是统计可靠的.

首先, 考虑平稳性的问题. 仿真的输出结果是仿真时间长度的函数. 初始的阶段会得到错误的行为, 这个阶段通常称为**瞬态的** (transient) 或者**预热阶段** (warm-up period). 当输出的结果变得平稳时, 系统运行进入**稳态** (steady state), 不幸的是, 没有办法预先估计从开始到稳态需要多长时间. 一般地, 仿真执行的时间越长, 达到稳态的机会就越大, 这一点在例 16.1-1 中也有说明, 即采用蒙特卡罗方法估计圆的面积的准确性随着样本数量的增大而提高. 因此, 可以通过使用足够大的样本数量来处理非平稳性.

其次, 我们考虑仿真观测需要从正态总体中采样的必要性. 可以通过使用*中心极限定理*(见 14.4.4 节) 来认识这种需要, 中心极限定理指出不管样本母体的分布如何, 这个抽样的平均意义的分布趋向于正态分布. 因此, 中心极限定理是我们用来满足正态分布假设的主要工具.

第 3 个条件涉及观测的独立性. 仿真实验的性质不能够保证连续的仿真观测之间是独立的. 然而, 通过使用样本平均来表示一个仿真观测, 我们可以减轻缺乏独立性带来的问题. 尤其是当我们增大用于求样本平均的仿真时间时它更加有效.

在讨论并解决了仿真实验中的特性之后, 我们给出 3 种常用的仿真观测数据收集的方法:

(1) 子区间法; (2) 重复实验方法; (3) 再生 (或者循环) 方法.

16.6.1 子区间法

图 16.10 刻画了子区间法的基本思想. 假定仿真执行 T 个时间单位 (即运行长度 $=T$), 并且需要得到 n 个观测值. 子区间法首先截去初始的瞬态期, 然后将剩余的仿真运行时间等分为 n 个子区间 (或者批次). 那么每个子区间上需要的平均测量值 (比如队列长度或者在队列中的等待时间) 可以用来代表单个的观测值. 将初始的瞬态期截去就意味着不会从这个阶段收集统计数据.

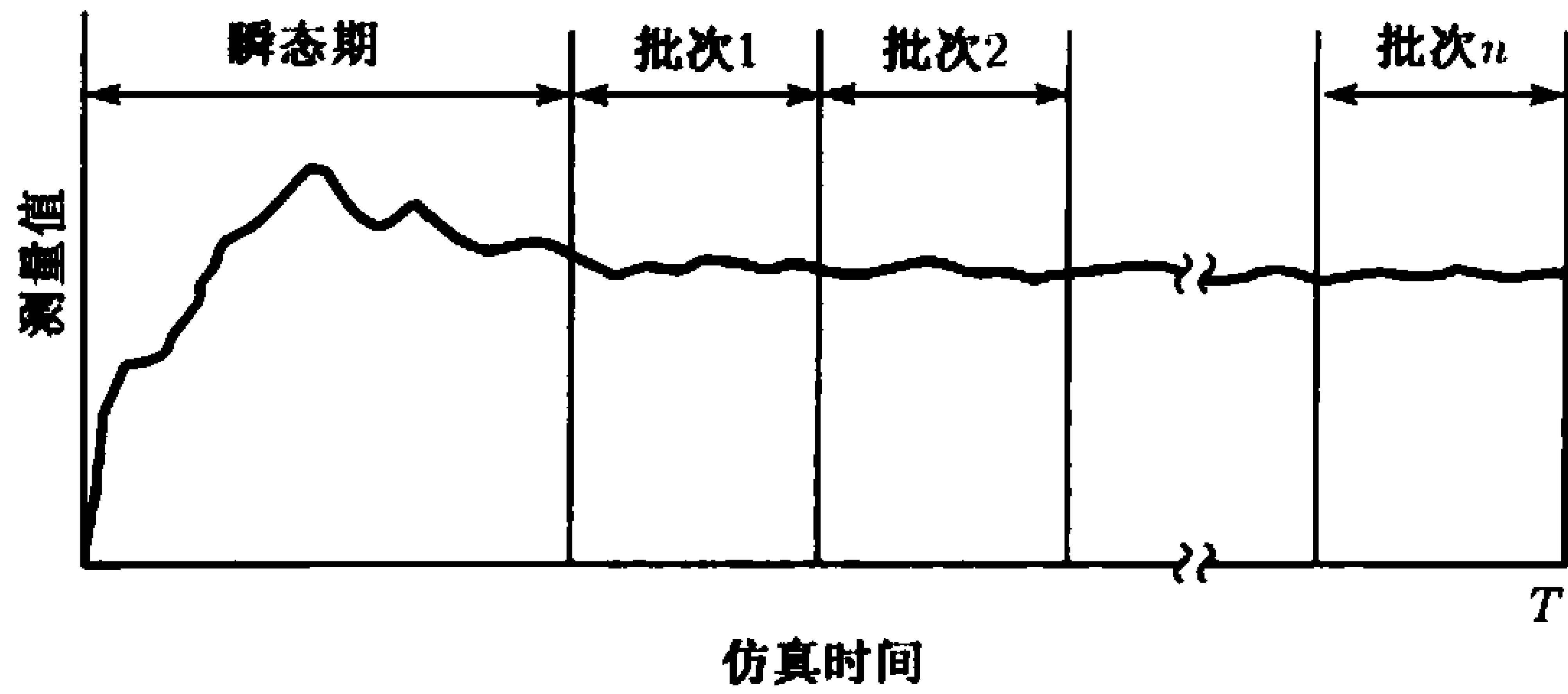


图 16.10 使用子区间法收集仿真数据

子区间法的优点就是减弱了仿真中瞬态的 (非稳态的) 条件带来的影响, 尤其是那些在仿真实验将要结束时的观测值. 子区间法的缺点是有相同边界条件的连续的两个子区间会是相关的, 当然可以通过增大每个批次的时间来削弱这种相关性.

例 16.6-1

图 16.11 显示了在单服务台排队模型中队列长度是仿真时间的函数的变化情况. 仿真实验长度为 $T = 35$ 小时, 瞬态期的长度估计为 5 小时. 现在需要收集 5 个观测值, 即 $n = 5$. 相应每个批次的时间长度为 $\frac{35-5}{5} = 6$ 小时.

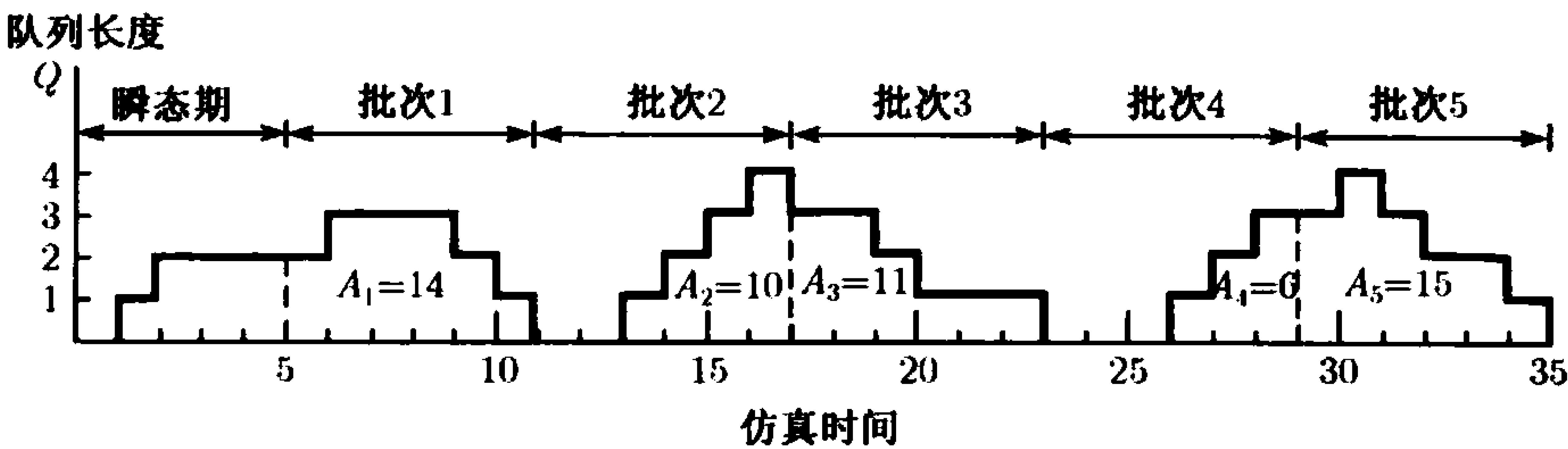


图 16.11 例 16.6-1 中队列长度随仿真时间变化的情况

用 $\overline{Q}_i$ 表示批次 i 的队列平均长度. 因为队列长度是基于时间的变量, 因此

$$\overline{Q}_i = \frac{A_i}{t}, i = 1, 2, \cdots, 5$$

其中 A_i 表示相应于批次 i 的队列长度曲线下方的面积, t 表示每个批次的时间, 在本例题中 $t = 6$ 小时.

根据图 16.11 中的数据, 可以得到下面的观测值:

观测值 i	1	2	3	4	5
A_i	14	10	11	6	15
$\overline{Q}_i$	2.33	1.67	1.83	1.00	2.5
样本均值 =1.87			样本标准差 =0.59		

如果需要的话, 样本均值和标准差可以用来计算置信区间. 例 16.6-1 中样本标准差的计算公式是

$$s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1}}$$

由于上面的式子忽略了相邻两个批次之间的相关性, 所以它只是真实标准差的一个近似值. 更加精确的计算公式可以参见 Law and Kelton(2000, pp. 249-253).

16.6.2 重复实验方法

正如图 16.12 所示, 在重复实验方法中, 每次观测都由一个截去瞬态期的独立

仿真实验所表示. 每一个批次的平均观测值的计算方法与子区间法的计算方法相同. 因为批次之间是不相关的, 所以唯一的差异就是可以使用标准差的计算公式.

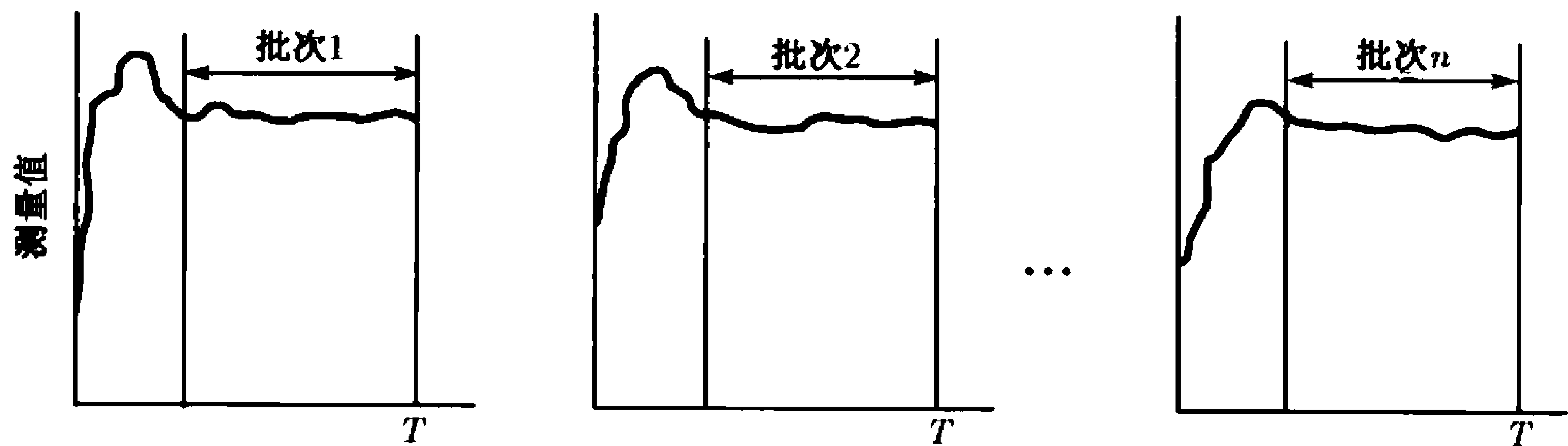


图 16.12 使用重复实验方法收集仿真数据

由于重复仿真实验方法每次运行都是使用不同的 (0,1) 随机数序列, 这就使得观测值具有真正的统计独立性, 这也正是重复实验方法的优点所在. 这种方法的缺点是初始的瞬态条件会带来相应的统计偏差, 可以通过充分延长仿真运行时间来减小它.

16.6.3 再生 (循环) 方法

再生方法可以视为子区间法的扩展, 这种新方法的出发点是为了减小自相关性带来的影响, 自相关性来源于子区间法要求每一个批次都有相同的初始条件. 例如, 如果我们现在考虑队列长度变量, 那么每一个批次都从队列长度为 0 开始. 与子区间法不同的是, 再生方法的特性会导致不同的批次有不同的时间长度.

虽然再生方法可以减轻自相关性的影响, 但是它存在的缺点是, 如果给定总的仿真运行长度, 那么批次的数目比较小, 这是因为我们无法预计一个新的批次什么时候开始以及这个批次要持续多长时间. 在稳态条件下, 我们期望每个批次的开始时间都是均匀地间隔开.

在再生方法中, 批次 i 的均值计算方法一般定义为两个随机变量 a_i 和 b_i 的比值, 即 $x_i = \frac{a_i}{b_i}$, 其中 a_i 和 b_i 的定义依赖于所要计算的变量. 具体来讲, 如果变量是基于时间的, 那么 a_i 表示曲线下方的面积, b_i 等于对应的总时间. 如果变量是基于观测的, 那么 a_i 等于批次 i 内的观测值总和, b_i 等于相应观测值的数目.

因为 x_i 是两个随机变量的比, 所以样本平均的一个无偏估计可以表示为

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

其中,

$$y_i = \frac{n\overline{a}}{\overline{b}} - \frac{(n-1)(n\overline{a} - a_i)}{n\overline{b} - b_i}, \quad i = 1, 2, \cdots, n$$

$$\overline{a}=\frac{\sum_{i=1}^{n} a_i}{n}$$

$$\overline{b}=\frac{\sum_{i=1}^{n} b_i}{n}$$

在这种情况下, 置信区间依赖于 y_i 的均值和标准差.

例 16.6-2

图 16.13 表示了在有 3 个并行服务台的单个设备中的忙碌的服务台数目. 仿真实验时间长度为 35 个时间单位, 瞬态期的时间长度为 4 个时间单位. 使用再生方法估计设备的平均使用率.

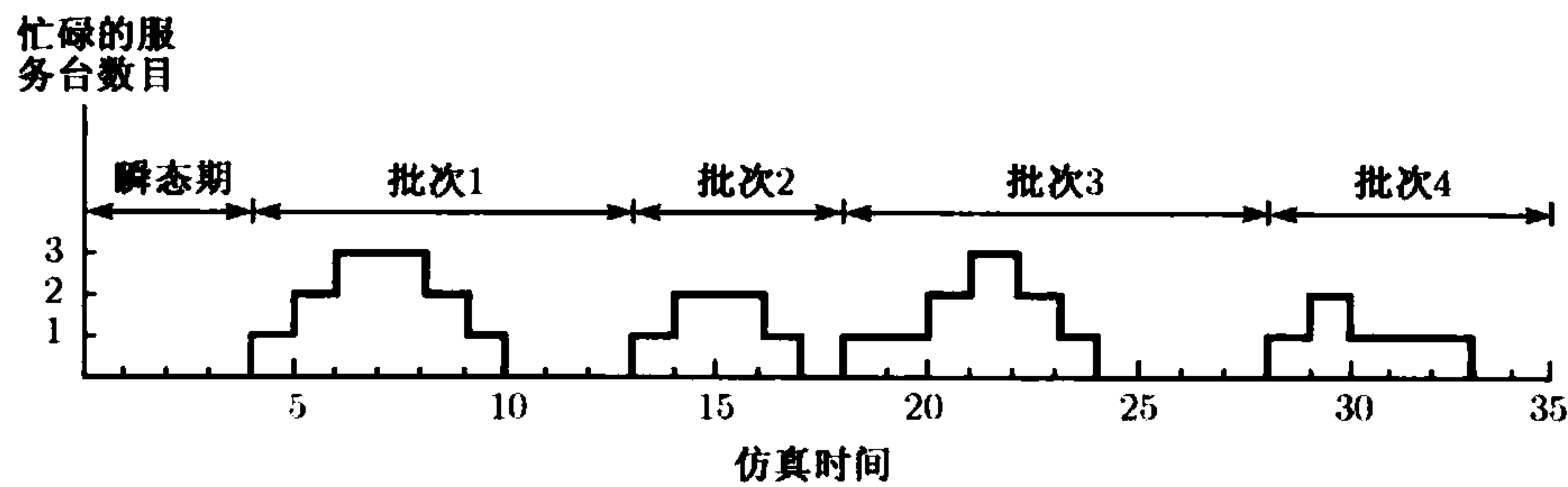


图 16.13 例 16.6-2 中忙碌的服务台数目作为仿真时间函数的变化情况

在截去瞬态期之后, 图 16.13 给出了 4 个批次, 每个批次都是从 3 个服务台全为空闲时开始. 下表给出了相应的 a_i 和 b_i 的值:

批次 i	a_i	b_i
1	12	9
2	6	5
3	10	10
4	6	7
均值	$\overline{a}=8.50$	$\overline{b}=7.75$

根据这些数据可以得到:

$$y_i=\frac{4\times 8.5}{7.75}-\frac{(4-1)\times(4\times 8.5-a_i)}{4\times 7.75-b_i}=4.39-\frac{102-3a_i}{31-b_i}$$

这些计算可以使用 Excel 模板 excelRegenerative.xls 更自动地计算.

习题 16.6A

1. 在例 16.6-1 中, 使用子区间法计算那些必须在队列中等待的顾客的平均等待时间.

*2. 在仿真模型中, 子区间法用于计算批次的均值, 瞬态期估计为 100, 每一个批次也刚好是 100 个时间单位. 使用下面的数据, 它们提供了顾客等待时间与仿真时间的关系, 估计平均等待时间的 95% 的置信区间.

时间区间	等待时间
0～100	10, 20, 13, 14, 8, 15, 6, 8
100～200	12, 30, 10, 14, 16
200～300	15, 17, 20, 22
300～400	10, 20, 30, 15, 25, 31
400～500	15, 17, 20, 14, 13
500～600	25, 30, 15

3. 在例 16.6-2 中, 假定每一个批次的开始时间都是在所有服务台恰好空闲的时刻. 那么, 在图 16.13 中, 这些时刻对应在 $t = 10, 17, 24, 33$. 根据再生点新的定义, 计算服务台利用率的 95% 的置信区间.

4. 在单个服务台排队模型中, 假定仿真系统实验时间为 100 小时, 仿真的结果显示只有在下面的几个时间区间上服务台才是忙碌的: (0,10),(15,20),(25,30),(35,60),(70,80),(90,95). 瞬态期的时间长度估计为 10 个小时.
   (a) 定义执行再生方法时需要的观测起始点.
   (b) 运用再生方法, 计算服务台平均利用率的 95% 的置信区间.
   (c) 对于同一个问题, 假定样本数量为 $n = 5$, 应用子区间法. 计算相应的 95% 的置信区间, 并且将这个结果与使用再生方法得到的结果进行比较.

16.7 仿真语言

执行仿真模型需要两种不同类型的运算: (1) 处理模型事件的排序存储和加工的文件操作; (2) 有关随机样本生成和模型统计数据收集的算术运算和记录. 第 1 种类型的运算包括链表加工中广义的逻辑运算, 第 2 种类型需要繁琐和耗时的计算. 由于这些计算的特点, 计算机成为完成仿真模型计算的必要工具, 因而也就促进了特殊的计算机仿真语言的发展, 以便于能够更方便、更有效地完成这些计算.

可用的离散仿真语言分为两大类:

(1) 事件时间表类; (2) 过程模拟类.

在事件时间表类语言中, 用户详细描述与每个事件的发生相关的行为, 与例 16.5.1 节中的过程大致相同. 这种情形下语言的主要作用是: (1) 从分布中自动抽样; (2) 按发生时间顺序存储和取出事件; (3) 收集模型统计数据.

过程模拟类的语言则是使用可以相互连接形成一个网络的块或者点, 来刻画**事务** (transaction) 或者**实体** (entity)(即, 顾客) 在整个系统中的移动情况. 例如, 在任何一个过程仿真语言中都有 3 个显著的块/点: 一个**源点**来产生事务, 一个**队列**用于放置需要排队等候的顾客, 一个**设备**用于给顾客提供服务. 定义每一个块/点, 其中包含可以使得仿真自动执行所需要的信息. 例如, 一旦指定了源点产生顾客的时间间隔, 过程模拟类的语言就自动知道了顾客到达事件什么时候发生. 实际上, 模

型中的每一个块/点都包含了明确的仿真网络中实体如何移动以及什么时间移动的指令.

与事件时间表类语言一样, 过程模拟类的语言也是通过同样的行为内部驱动的. 不同的是, 这些行动都是自动地减轻用户繁琐的计算和逻辑细节. 在某种程度上, 可以把过程模拟类的语言认为是基于 "黑箱" 办法的输入–输出思想. 这实质上意味着过程模拟的语言倾向于简单易用.

事件时间表的语言主要有 SIMSCRIPT、SLAM 和 SIMAN. 经过多年的发展, 这些语言已经具有过程模拟的能力. 这 3 种语言都可以允许用户在高级语言环境中写 (部分) 模型, 例如 FORTRAN 或者 C. 正是这种能力使得它们可以模拟更复杂的情形, 而这些情形不能使用原始的仿真语言直接模拟. 这种局限性来源于这些语言在模型队列和设施之间移动事务 (或者实体) 所使用的受限制的或者循环的规则.

如今仿真市场上几个主要的商用软件包包括 Arena、AweSim 和 GPSS/H, 这些软件包使用扩展的用户界面来简化建立仿真模型的过程. 它们还提供动画功能, 使得系统的变化成为可视化的. 然而, 对于一些有经验的用户来说, 这种界面反而会制约开发仿真模型的速度, 因此某些用户更倾向于使用传统的程序语言来开发仿真模型, 如 C、Basic 和 FORTRAN.

习题 16.7A[①]

1. 顾客随机到达一个有 3 个办事员的邮局. 到达时间间隔服从期望是 5 分钟的指数分布. 一个办事员为一个顾客服务的时间服从期望为 10 分钟的指数分布, 所有到达的顾客排成一个队列, 等待空闲的办事员. 运行这个系统的仿真模型 480 分钟, 然后确定下面的各个指标:
   (a) 顾客在队列中的平均等待时间.
   (b) 办事员的平均利用率.
   (c) 将得到的结果分别与 $M/M/c$ 排队模型 (见第 12 章) 的结果和 MultiServerSimulator.xls 电子表格的结果进行比较.
2. 电视元件在传送带上以每小时 5 个的均匀速度到达并接受检查, 检查时间服从 10~15 分钟的均匀分布. 根据以往的经验知道, 有 20% 的受检查元件会被调整并重新接受检查, 调整的时间服从 6~8 分钟的均匀分布. 运行这个仿真模型 480 分钟, 然后确定下面的各个指标:
   (a) 一个元件通过检查的平均时间.
   (b) 一个元件在离开这个系统之前需要接受重复检查的平均次数.
3. 小鼠被困在迷宫中但拼命 "想出来". 经过尝试一段服从 1~3 分钟的均匀分布的时间之后, 它有 30% 的机会可以找到正确的道路, 否则它将无目的的徘徊一段服从 2~3 分钟的均匀分布的时间, 然后回到它开始尝试的地方, 并进行下一次尝试. 当然, 小鼠可以尝试任意多的次数, 但是任何事情总有结束, 随着小鼠在尝试再尝试中体力的消耗, 如果在经过一段服

① 利用自己选择的仿真语言完成下面的习题, 或者使用 BASIC、FORTRAN 或 C.

从期望是 10 分钟, 标准差为 2 分钟的正态分布的时间之后, 它仍然没有找到出口的话将死去. 建立一个仿真模型来估计小鼠可以获得自由的概率. 为了估计这个概率, 假定模型使用 100 只小鼠.

4. 在汽车制造的最后阶段, 汽车在两侧各有一个工作站的传送带上传送, 这样可以同时对汽车的两侧进行加工. 左侧和右侧的加工时间分别服从 15~20 分钟和 18~22 分钟的均匀分布, 汽车在传送带上每 20 到达一辆. 仿真这个过程 480 分钟, 来确定左侧和右侧工作站的利用率.

5. 在一个每次只能洗一辆车的洗车公司, 汽车到达的时间间隔服从期望为 10 分钟的指数分布. 车辆到达以后排成一个队列, 队列最多只能有 5 辆车等候. 如果队列已满, 那么新来的车不得不去其他的洗车公司. 清洗一辆车的时间服从 10~15 分钟的均匀分布. 仿真这个系统 960 分钟, 并估计一辆车需要在洗车间的平均时间.

参考文献

Box, G., and M. Muller, "A Note on the Generation of Random Normal Deviates," *Annals of Mathematical Statistics*, Vol. 29, pp. 610-611, 1958.

Law, A., and W. Kelton, *Simulation Modeling & Analysis*, 3rd ed., McGraw-Hill, New York, 2000.

Ross, S., *A Course in Simulation*, Macmillan, New York, 1990.

Rubenstein, R., B. Melamed, and A. Shapiro, *Modern Simulation and Modeling*, Wiley, New York, 1998.

Taha, H., *Simulation Modeling and SIMNET*, Prentice Hall, Upper Saddle River, NJ, 1988.

第 17 章　马尔可夫链

本章导读　本章介绍马尔可夫链的基本背景及其实际应用，包括基于费用的模型. 马尔可夫链的记法很“麻烦”，计算上也比较繁琐. 针对这一问题，我们尽可能采用相对简单的矩阵记法. 在计算方面，我们提供了两个 Excel 程序，用于作任意大小的马尔可夫链的基本计算，包括求 n 步转移和绝对概率值、平稳状态概率以及计算遍历以及吸收链的首次通过时间. 这两个电子表格程序都可以用于求解节后习题.

本章包括 17 个例题、2 个电子表格程序和 47 个节后习题. AMPL/Excel/Solver/TORA 程序放在文件夹 ch17Files 下.

17.1　马尔可夫链的定义

令 X_t 为一个随机变量，表示在时间 $t = 1, 2, \cdots$ 离散点上的系统状态. 这个随机变量族 $\{X_t\}$ 就形成了一个**随机过程** (stochastic process). 一个随机过程的状态数可以是有限的，也可以是无限的，如下面两个例子所示.

例 17.1-1 (机器维护)

一台机器在每月的保养维护时的状态条件用较差、一般或良好来表征. 对第 t 个月，这个问题的随机过程可以表示为

$$X_t = \left\{ \begin{array}{ll} 0, & \text{若机器条件较差} \\ 1, & \text{若机器条件一般} \\ 2, & \text{若机器条件良好} \end{array} \right\}, t = 1, 2, \cdots$$

该随机变量 X_t 是有限的，因为它只表示 3 个状态：较差 (0)，一般 (1)，良好 (2).

例 17.1-2 (机件加工车间)

工件按平均每小时 5 件的速率到达某机件加工车间. 到达过程服从泊松分布：理论上，在时间区间 $(0, t)$ 内，它允许 0 到无穷大之间任何数量的工件来到该车间. 这一描述到来工件的无限状态的过程为：

$$X_t = 0, 1, 2, \cdots, \quad t > 0$$

一个随机过程是**马尔可夫过程** (Markov process)，若某个未来状态的发生仅仅依赖于前一个状态. 这就意味着，对给定的时间顺序 $t_0, t_1, \cdots, t_n$，若随机变量族 $\{X_t\}$ 具有以下性质，则称它为一个马尔可夫过程：

$$P\{X_{t_n}=x_n|X_{t_{n-1}}=x_{n-1},\cdots,X_{t_0}=x_0\}=P\{X_{t_n}=x_n|X_{t_{n-1}}=x_{n-1}\}$$

在一个具有 n 个完全的和互斥状态 (结果) 的马尔可夫过程中, 在时间的特定点 $t=0,1,2,\cdots$ 上的概率通常记为

$$p_{ij}=P\{X_t=j|X_{t-1}=i\},\quad i,j=1,2,\cdots,n,\ t=0,1,2,\cdots,T$$

上式称为从 $t-1$ 时的状态 i 移动到 t 时的状态 j 的**一步转移概率** (one-step transition probability). 按照定义, 我们有

$$\sum_j p_{ij}=1,\quad i=1,2,\cdots,n$$
$$p_{ij}\geqslant 0,\quad i,\ j=1,2,\cdots,n$$

为了方便描述一步转移概率, 采用下面的矩阵记法:

$$\boldsymbol{P}=\begin{pmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1n}\\ p_{21} & p_{22} & p_{23} & \cdots & p_{2n}\\ \vdots & \vdots & \vdots & & \vdots\\ p_{n1} & p_{n2} & p_{n3} & \cdots & p_{nn}\end{pmatrix}$$

矩阵 $\boldsymbol{P}$ 就定义了我们所称的**马尔可夫链** (Markov chain), 其性质是, 它的所有转换概率 p_{ij} 都是固定的 (稳定的), 而且与时间无关. 虽然马尔可夫链可以包括无穷多个状态, 本章只介绍有限多个链, 因为本章只涉及这一类情形.

例 17.1-3 (园艺师问题)

每年园艺季节 (4 月到 9 月) 开始的时候, 有一位园丁要对土壤条件进行化学试验. 根据试验结果, 把新园艺季节的土壤肥力划分为下面的 3 种状态之一: (1) 良好; (2) 一般; (3) 较差. 这位园艺师通过多年的观察发现, 上一年的土壤条件会影响到当前年的肥力, 而且这种情况可以用下面的马尔可夫链来描述:

$$\boldsymbol{P}=\text{今年系统的状态}\begin{cases}1\\2\\3\end{cases}\overset{\substack{\text{下一年系统的状态}\\ 1\quad\ \ 2\quad\ \ 3}}{\begin{pmatrix}0.2 & 0.5 & 0.3\\ 0 & 0.5 & 0.5\\ 0 & 0 & 1\end{pmatrix}}$$

这些转移概率表明, 土壤条件要么退化了, 要么保持不变, 但不会改善. 如果今年的土壤条件良好 (状态 1), 则有 20% 的机会使得明年不会改变, 有 50% 的机会明年会变得一般 (状态 2), 而有 30% 的机会使得明年会退化成较差的条件 (状态 3). 假如今年的土壤条件一般 (状态 2), 则明年的土壤肥力也是一般的概率为 0.5,

变成较差条件的概率也是 0.5. 最后, 今年较差的土壤条件 (状态 3) 则只能导致明年同样的土壤条件 (概率为 1).

该园艺师可以通过使用肥料来改变转换概率 $\boldsymbol{P}$, 即提高土壤肥力. 在这种情况下, 转换矩阵就变成了:

$$\boldsymbol{P}_1 = \begin{matrix} \\ 1 \\ 2 \\ 3 \end{matrix} \begin{matrix} \begin{matrix} 1 & \quad 2 & \quad 3 \end{matrix} \\ \begin{pmatrix} 0.30 & 0.60 & 0.10 \\ 0.10 & 0.60 & 0.30 \\ 0.05 & 0.40 & 0.55 \end{pmatrix} \end{matrix}$$

现在通过使用化肥可以让退化的土壤条件得以改善, 有 10% 的机会能让土壤条件从一般变成良好 (从状态 2 变成状态 1), 有 5% 的机会能让土壤条件从较差变成良好 (从状态 3 变成状态 1), 并且有 40% 的机会能让土壤条件从较差改变成一般 (从状态 3 变成状态 2).

习题 17.1A

1. 一位工程学教授每两年就要购买一台新的电脑, 他从 3 种型号中选购: M1、M2 和 M3. 如果现在的电脑型号是 M1 的话, 下一台电脑为 M2 型的概率为 0.2, 购买 M3 型的概率是 0.15. 假如他现在的电脑型号是 M2 的话, 下一台电脑为 M1 和 M3 型的概率分别为 0.6 和 0.25. 如果他现在的电脑型号是 M3 的话, 则换成 M1 和 M2 型电脑的概率分别为 0.5 和 0.1. 用马尔可夫链来表示这一情形.

***2.** 一辆警车在一个已知有坏蛋的社区巡逻. 在巡逻过程中, 有 60% 的机会能按时到达求助现场, 其他情况下警车将继续正常巡逻. 在收到求助请求后, 有 10% 的机会会取消 (在这种情况下可恢复正常巡逻), 有 30% 的机会该警车仍在响应前一个请求. 当警车到达现场时, 有 10% 的机会嫌犯已经跑掉了 (在这种情况下警车返回巡逻任务), 而有 40% 的机会会立即对嫌疑犯进行拘捕, 其他情况下, 警察会对该区域进行搜索. 假如发生拘捕情况, 有 60% 的机会会把嫌疑犯送到警察局, 其他情况下他们会放了嫌疑人, 警车返回巡逻任务. 请用转移矩阵表示警察巡逻的概率行为.

3. (Cyert 等人, 1963) 某银行提供贷款, 这些贷款有按期偿还的, 也有迟期还款的. 若一笔贷款的还款晚了 4 个季度 (一年) 以上, 银行就认为该贷款是坏账, 将其注销. 下表给出了该银行贷款的历史纪录.

贷款额度	迟还款季度	还款历史纪录
$10 000	0	正常还款 $2 000, 还 $3 000 晚了 1 个季度, 还 $3 000 晚了 2 个季度, 其他晚了 3 个季度
$25 000	1	正常还款 $4 000, 还 $12 000 晚了 1 个季度, 还 $6 000 晚了 2 个季度, 其他晚了 3 个季度
$50 000	2	正常还款 $7 500, 还 $15 000 晚了 1 个季度, 其他还款晚了 2 个季度
$50 000	3	正常还款 $42 000, 其他还款晚了 1 个季度
$100 000	4	正常还款 $50 000

将该银行的贷款情形表示为一个马尔可夫链.

4. (Pliskin and Tell, 1981) 对患有肾脏病的病人可采用肾移植手术, 或者进行定期透析治疗. 在任何一年中, 有 30% 的患者进行了尸体肾移植, 有 10% 的人接受了活体捐赠的肾移植. 在进行了肾移植手术的第 2 年, 有 30% 的尸体肾移植和 15% 的活体捐助肾移植接受者需要重新进行透析治疗. 这两组病人的死亡率分别为 20% 和 10%. 在透析人群中, 有 10% 的人死亡了, 在接受肾移植后存活 1 年以上的人中, 死亡率为 5%, 而且有 5% 需要重新进行肾透析. 请用马尔可夫链表示上述情形.

17.2 绝对转移概率和 n 步转移概率

给定状态 j 开始时的初始概率 $\boldsymbol{a}^{(0)} = \{a_j^{(0)}\}$ 以及一个马尔可夫链的转移矩阵 $\boldsymbol{P}$, 在 n 次转移 $(n > 0)$ 后处于状态 j 的绝对概率按如下计算:

$$\begin{aligned}
\boldsymbol{a}^{(1)} &= \boldsymbol{a}^{(0)}\boldsymbol{P} \\
\boldsymbol{a}^{(2)} &= \boldsymbol{a}^{(1)}\boldsymbol{P} = \boldsymbol{a}^{(0)}\boldsymbol{P}\boldsymbol{P} = \boldsymbol{a}^{(0)}\boldsymbol{P}^2 \\
\boldsymbol{a}^{(3)} &= \boldsymbol{a}^{(2)}\boldsymbol{P} = \boldsymbol{a}^{(0)}\boldsymbol{P}^2\boldsymbol{P} = \boldsymbol{a}^{(0)}\boldsymbol{P}^3
\end{aligned}$$

继续同样的步骤, 我们得到

$$\boldsymbol{a}^{(n)} = \boldsymbol{a}^{(0)}\boldsymbol{P}^n, n = 1, 2, \cdots$$

矩阵 $\boldsymbol{P}^n$ 称为**n 步转移矩阵** (n-step transition matrix). 从这些计算中我们可以看出

$$\boldsymbol{P}^n = \boldsymbol{P}^{n-1}\boldsymbol{P}$$

或

$$\boldsymbol{P}^n = \boldsymbol{P}^{n-m}\boldsymbol{P}^m, \quad 0 < m < n$$

这些称为**Chapman-Kolomogorov(查普曼-科尔莫戈罗夫) 方程**.

例 17.2-1

下面的转移矩阵应用于施肥情况下的园艺师问题 (例 17.1-3):

$$\boldsymbol{P} = \begin{matrix} & \begin{matrix} 1 & \;\;\;2 & \;\;\;3 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} & \begin{pmatrix} 0.30 & 0.60 & 0.10 \\ 0.10 & 0.60 & 0.30 \\ 0.05 & 0.40 & 0.55 \end{pmatrix} \end{matrix}$$

土壤的初始条件为良好, 即 $\boldsymbol{a}^{(0)} = (1, 0, 0)$. 求 1 个、8 个和 16 个园艺季节以后系

统 3 种状态的绝对概率.

$$P^2=\begin{pmatrix}0.30 & 0.60 & 0.10\\ 0.10 & 0.60 & 0.30\\ 0.05 & 0.40 & 0.55\end{pmatrix}\begin{pmatrix}0.30 & 0.60 & 0.10\\ 0.10 & 0.60 & 0.30\\ 0.05 & 0.40 & 0.55\end{pmatrix}=\begin{pmatrix}0.155\,0 & 0.580\,0 & 0.265\,0\\ 0.105\,0 & 0.540\,0 & 0.355\,0\\ 0.082\,5 & 0.490\,0 & 0.427\,5\end{pmatrix}$$

$$P^4=\begin{pmatrix}0.155\,0 & 0.580\,0 & 0.265\,0\\ 0.105\,0 & 0.540\,0 & 0.355\,0\\ 0.082\,5 & 0.490\,0 & 0.427\,5\end{pmatrix}\begin{pmatrix}0.155\,0 & 0.580\,0 & 0.265\,0\\ 0.105\,0 & 0.540\,0 & 0.355\,0\\ 0.082\,5 & 0.490\,0 & 0.427\,5\end{pmatrix}$$

$$=\begin{pmatrix}0.106\,79 & 0.532\,95 & 0.360\,26\\ 0.102\,26 & 0.526\,45 & 0.371\,29\\ 0.099\,50 & 0.521\,93 & 0.378\,57\end{pmatrix}$$

$$P^8=\begin{pmatrix}0.106\,79 & 0.532\,95 & 0.360\,26\\ 0.102\,26 & 0.526\,45 & 0.371\,29\\ 0.099\,50 & 0.521\,93 & 0.378\,57\end{pmatrix}\begin{pmatrix}0.106\,79 & 0.532\,95 & 0.360\,26\\ 0.102\,26 & 0.526\,45 & 0.371\,29\\ 0.099\,50 & 0.521\,93 & 0.378\,57\end{pmatrix}$$

$$=\begin{pmatrix}0.101\,753 & 0.525\,514 & 0.372\,733\\ 0.101\,702 & 0.525\,435 & 0.372\,863\\ 0.101\,669 & 0.525\,384 & 0.372\,863\end{pmatrix}$$

$$P^{16}=\begin{pmatrix}0.101\,753 & 0.525\,514 & 0.372\,733\\ 0.101\,702 & 0.525\,435 & 0.372\,863\\ 0.101\,669 & 0.525\,384 & 0.372\,863\end{pmatrix}\begin{pmatrix}0.101\,753 & 0.525\,514 & 0.372\,733\\ 0.101\,702 & 0.525\,435 & 0.372\,863\\ 0.101\,669 & 0.525\,384 & 0.372\,863\end{pmatrix}$$

$$=\begin{pmatrix}0.101\,659 & 0.524\,54 & 0.372\,881\\ 0.101\,659 & 0.524\,54 & 0.372\,881\\ 0.101\,659 & 0.524\,54 & 0.372\,881\end{pmatrix}$$

因此有

$$a^{(1)}=(1\ \ 0\ \ 0)\begin{pmatrix}0.30 & 0.60 & 0.10\\ 0.10 & 0.60 & 0.30\\ 0.05 & 0.40 & 0.55\end{pmatrix}=(0.30\ \ 0.60\ \ 0.1)$$

$$a^{(8)}=(1\ \ 0\ \ 0)\begin{pmatrix}0.101\,753 & 0.525\,514 & 0.372\,733\\ 0.101\,702 & 0.525\,435 & 0.372\,863\\ 0.101\,669 & 0.525\,384 & 0.372\,863\end{pmatrix}=(0.101\,753\ \ 0.525\,514\ \ 0.372\,733)$$

$$a^{(16)} = (1\ \ 0\ \ 0)\begin{pmatrix} 0.101\,659 & 0.524\,54 & 0.372\,881 \\ 0.101\,659 & 0.524\,54 & 0.372\,881 \\ 0.101\,659 & 0.524\,54 & 0.372\,881 \end{pmatrix}$$
$$= (0.101\,659\ \ 0.524\,54\ \ 0.372\,881)$$

P^8 的各行与绝对概率向量 $a^{(8)}$ 几乎完全相同, P^{16} 的情形更是如此. 这说明, 随着转移次数的增加, 绝对概率与初始的 $a^{(0)}$ 无关. 在这种情况下, 所得到的概率称为**稳态概率** (steady-state probabilities).

评注 关于马尔可夫链的计算相当繁琐. excelMarkovChains.xls 是一个方便的通用电子表格程序, 可用来进行这些计算 (见后面例 17.4-1 的*Excel* 程序).

习题 17.2A

1. 考虑习题 17.1A 中的第 1 题, 求这位教授在未来 4 年中将会购买当前型号电脑的概率.

*2. 考虑习题 17.1A 中的第 2 题, 假如警车现正在求助现场, 求在两次巡逻中会发生拘捕的概率.

3. 考虑习题 17.1A 中的第 3 题. 假设该银行现在有价值 \$500 000 的未偿还贷款, 其中 \$100 000 是新贷款, \$50 000 拖延 1 个季度偿还, \$150 000 拖延了 2 个季度, \$100 000 拖延了 3 个季度, 其余的晚了 4 个季度. 两个贷款周期以后, 这些贷款的情形会是如何?

4. 考虑习题 17.1A 中的第 4 题.
   (a) 对于一个现在处于透析治疗的病人, 在未来 2 年内接受肾移植的概率是多少?
   (b) 对于一个现在存活了 1 年以上的病人, 再存活 4 年的概率是多少?

17.3 马尔可夫链中状态的分类

一个马尔可夫链的状态可以按照 P 的转移概率 p_{ij} 进行分类.

(1) 状态 j 是**吸收的** (absorbing), 若它经过一次转移肯定会返回到它自身, 即 $p_{jj} = 1$.

(2) 状态 j 是**瞬时的** (transient), 若它能达到另一种状态, 但不能从另一种状态回到自身. 在数学上, 若对所有的 i, $\lim\limits_{n\to\infty} p_{ij}^{(n)} = 0$, 则会出现这种情况.

(3) 状态 j 是**循环的** (recurrent), 若从其他状态返回的概率为 1. 出现这种情况, 当且仅当该状态为非瞬时的.

(4) 状态 j 是**周期性的** (periodic), 周期为 $t > 1$, 若仅在 $t, 2t, 3t, \cdots$ 步可能出现返回. 这就意味着, 当 n 不能被 t 整除时, $p_{jj}^{(n)} = 0$.

根据上述定义, 一个*有限的*马尔可夫链不能只含有瞬时的状态, 因为按照定义, 瞬时性要求进入其他的"陷阱"状态, 因此不会返回到瞬时状态. 这种"陷阱"状态不一定是单一的吸收状态. 例如在链

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0.3 & 0.7 \\ 0 & 0 & 0.4 & 0.6 \end{pmatrix}$$

中, 状态 1 和状态 2 是瞬时的, 因为一旦系统被"陷入"状态 3 和状态 4, 它们就不能再进来了. 在这种情况下, 状态 3 和状态 4 起到一种吸收状态的作用, 形成一种**闭集合** (closed set). 按照定义, 一个闭集合的所有状态必须**相通** (communicate), 这意味着有可能从任一个状态通过一步或多步转移到达这个集合的每个其他状态, 即对于所有的 $i \neq j$ 并且 $n \geqslant 1$, $p_{ij}^{(n)} > 0$. 注意到状态 3 和状态 4 都可以是吸收状态, 如果 $p_{33} = p_{44} = 1$. 在这种情况下, 每个状态都形成一个闭集合.

一个闭的马尔可夫链称为**各态遍历的** (ergodic), 如果它的所有状态都是循环的而且是非周期的(没有周期性). 在这种情况下, n 步转移后的绝对概率 $\boldsymbol{a}^{(n)} = \boldsymbol{a}^{(0)}\boldsymbol{P}^n$, 当 $n \to \infty$ 时, 总是唯一地收敛到一个与初始概率 $\boldsymbol{a}^{(0)}$ 无关的极限 (平稳状态) 分布, 我们将在 17.4 节进一步说明.

例 17.3-1 (吸收状态和瞬时状态)

考虑园艺师不用施肥情况下的马尔可夫链.

$$P = \begin{pmatrix} 0.2 & 0.5 & 0.3 \\ 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \end{pmatrix}$$

状态 1 和状态 2 是瞬时的, 因为它们达到状态 3 但永远不会返回. 状态 3 是吸收状态, 因为 $p_{33} = 1$. 通过计算 $\lim_{n\to\infty} p_{ij}^{(n)} = 0$, 也能看出这些分类结果. 例如

$$P^{(100)} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

这说明经过多次转移以后, 再进入瞬时状态 1 和状态 2 的概率为 0, 而"陷入"吸收状态 3 的概率为 1.

例 17.3-2 (周期状态)

通过计算 $\boldsymbol{P}^n$, 并对 $n = 2, 3, 4, \cdots$ 观察 $p_{ii}^{(n)}$ 的值, 可以检查一个状态的周期性. 这些值只有在状态的对应周期是正的. 例如在链

$$P = \begin{pmatrix} 0 & 0.6 & 0.4 \\ 0 & 1 & 0 \\ 0.6 & 0.4 & 0 \end{pmatrix}$$

中, 我们有

$$
\boldsymbol{P}^2=\begin{pmatrix}0.24 & 0.76 & 0\\ 0 & 1 & 0\\ 0 & 0.76 & 0.24\end{pmatrix},\quad \boldsymbol{P}^3=\begin{pmatrix}0 & 0.904 & 0.096\ 0\\ 0 & 1 & 0\\ 0.144 & 0.856 & 0\end{pmatrix},
$$

$$
\boldsymbol{P}^4=\begin{pmatrix}0.0576 & 0.942\,4 & 0\\ 0 & 1 & 0\\ 0 & 0.942\,4 & 0.057\,6\end{pmatrix},\quad \boldsymbol{P}^5=\begin{pmatrix}0 & 0.976\,96 & 0.023\,04\\ 0 & 1 & 0\\ 0.034\,56 & 0.965\,44 & 0\end{pmatrix}
$$

对于 $n=6,7,\cdots$, 继续计算 $\boldsymbol{P}^n$ 表明, 当 n 为偶数值时, p_{11} 和 p_{33} 为正的, 其他为零. 这说明状态 1 和状态 3 的周期为 2.

习题 17.3A

1. 对下列马尔可夫链的状态进行分类. 若状态是周期性的, 求其周期.

*(a) $\begin{pmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ 1 & 0 & 0\end{pmatrix}$

*(b) $\begin{pmatrix}\frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0\\ 0 & 0 & 1 & 0\\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3}\\ 0 & 0 & 0 & 1\end{pmatrix}$

(c) $\begin{pmatrix}0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0.5 & 0.5 & 0 & 0 & 0\\ 0 & 0.7 & 0.3 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 0.4 & 0.6\\ 0 & 0 & 0 & 0 & 0.2 & 0.8\end{pmatrix}$

(d) $\begin{pmatrix}0.1 & 0 & 0.9\\ 0.7 & 0.3 & 0\\ 0.2 & 0.7 & 0.1\end{pmatrix}$

17.4 遍历链的稳定状态概率和平均返回时间

在遍历型马尔可夫链中, 稳定状态概率定义为

$$
\pi_j=\lim_{n\to\infty} a_j^{(n)},\quad j=0,1,2,\cdots
$$

这些概率与 $\{a_j^{(0)}\}$ 无关, 可以从下面的公式求出:

$$
\boldsymbol{\pi}=\boldsymbol{\pi}\boldsymbol{P},\qquad \sum_j \pi_j=1
$$

(在 $\boldsymbol{\pi}=\boldsymbol{\pi}\boldsymbol{P}$ 中有一个方程是冗余的.) $\boldsymbol{\pi}=\boldsymbol{\pi}\boldsymbol{P}$ 说明在一次转移后, 概率 $\boldsymbol{\pi}$ 保持不变, 因此, 它们代表稳定状态的分布.

稳定状态概率的一个衍生结果是, 可以确定系统第一次返回到状态 j 之前的期望转移步数, 这被称为**平均首次返回时间** (mean first return time), 或称为**平均循**

环时间 (mean recurrence time), 并且在 n 个状态的马尔可夫链中按下式计算:

$$\mu_{jj} = \frac{1}{\pi_j}, \quad j = 1, 2, \cdots, n$$

例 17.4-1

为了求施肥情况下的园艺师问题 (例 17.1-3) 的稳定状态概率分布, 我们有

$$(\pi_1 \quad \pi_2 \quad \pi_3) = (\pi_1 \quad \pi_2 \quad \pi_3)\begin{pmatrix} 0.3 & 0.6 & 0.1 \\ 0.1 & 0.6 & 0.3 \\ 0.05 & 0.4 & 0.55 \end{pmatrix}$$

由此得出下面的一组方程:

$$\begin{aligned} \pi_1 &= 0.3\pi_1 + 0.1\pi_2 + 0.05\pi_3 \\ \pi_2 &= 0.6\pi_1 + 0.6\pi_2 + 0.4\pi_3 \\ \pi_3 &= 0.1\pi_1 + 0.3\pi_2 + 0.55\pi_3 \\ \pi_1 &+ \pi_2 + \pi_3 = 1 \end{aligned}$$

记住前 3 个方程中有一个 (任何一个) 是冗余的, 这组方程的解是 $\pi_1 = 0.101\,7$, $\pi_2 = 0.525\,4$, $\pi_3 = 0.372\,9$. 这些概率值说明, 经过多步转移后, 土壤条件大致为, 10% 的时候是良好的, 52% 的时候一般, 而 37% 的时候较差.

平均首次返回时间的计算为

$$\mu_{11} = \frac{1}{0.101\,7} = 9.83, \quad \mu_{22} = \frac{1}{0.525\,4} = 1.9, \quad \mu_{33} = \frac{1}{0.372\,9} = 2.68$$

这说明, 按照土壤当前的状态, 大约要花 10 个园艺季节使土壤返回到良好状态, 2 个园艺季节后返回一般状态, 3 个园艺季节返回到较差的土壤状态. 这些结果表明了按照所提出的施肥方案, 土壤条件改善的希望并不乐观, 需要施用更多的肥料才能改善这种状况. 例如考虑下面的转移矩阵, 其中转移到良好状态的概率比前面的矩阵要更大些:

$$\boldsymbol{P} = \begin{pmatrix} 0.35 & 0.6 & 0.05 \\ 0.3 & 0.6 & 0.1 \\ 0.25 & 0.4 & 0.35 \end{pmatrix}$$

在这种情况下, $\pi_1 = 0.31$, $\pi_2 = 0.58$, $\pi_3 = 0.11$, 得到 $\mu_{11} = 3.2$, $\mu_{22} = 1.7$, $\mu_{33} = 8.9$, 这与前面的"悲观"结果正好相反.

Excel 程序

图 17.1 显示了使用通用的 Excel 程序 excelMarkovChains.xls 对园艺师例子计算 n 步的绝对稳态概率, 以及计算任意大小马尔可夫链平均返回时间的输出结果.

计算步骤具有自我说明性. 在 2a 步中, 可以采用你自己的一套编码覆盖这些默认的状态编号 $(1,2,3,\cdots)$. 在执行第 4 步的时候, 这些编码会在电子表格的其他地方自动更新.

	A	B	C	D	E	F	G	H
1	Markov Chains							
2	Step 1:	Number of states =		3	Step 2a:	Initial probabilities:		
3	Step 2:				Codes:	1	2	3
4						1	0	0
5	Step 3:	Number of transitions=		8	Step 2b:	Input Markov chain		
6	Step 4:					1	2	3
7					1	0.3	0.6	0.1
8	Output Results				2	0.1	0.6	0.3
9		Absolute	Steady	Mean return	3	0.05	0.4	0.55
10	State	(8-step)	state	time		Output (8-step) transition matrix		
11	1	0.10175	0.101695	9.8333254		1	2	3
12	2	0.52551	0.525424	1.9032248	1	0.10175	0.525514	0.372733
13	3	0.37273	0.372882	2.6818168	2	0.1017	0.525435	0.372864
14					3	0.10167	0.525384	0.372947

图 17.1 计算马尔可夫链的 Excel 电子表格程序

例 17.4-2 (费用模型)

考虑施肥情况下的园艺师问题 (例 17.1-3). 假设肥料的费用为每袋 \$50, 并且如果土壤条件为良好, 那么花园需要 2 袋肥料. 若土壤条件一般, 则肥料的施用量增加 25%; 若土壤较差, 则增加 60%. 这位园艺师估计, 假如不使用肥料, 花园的年产值为 \$250; 若施用肥料的话, 年产值 \$420. 那么施肥值得吗?

利用例 17.4-1 中的稳态概率, 我们得到

$$\begin{aligned}
\text{肥料的年度期望费用} &= 2 \times \$50 \times \pi_1 + (1.25 \times 2) \times \$50 \times \pi_2 \\
&\quad + (1.60 \times 2) \times \$50 \times \pi_3 \\
&= 100 \times 0.101\,7 + 125 \times 0.525\,4 + 160 \times 0.372\,9 \\
&= \$135.51 \\
\text{增加年产值} &= \$420 - \$250 = \$170
\end{aligned}$$

这一结果表明, 平均来讲, 施用肥料的净收益为 $\$170 - \$135.51 = \$34.49$, 因此建议施肥.

习题 17.4A

*1. 在春季的某个晴朗日子里, MiniGolf 球场的收入是 \$2 000; 若是阴天的话, 则收入会下降 20%; 雨天的收入会减少 80%. 假如今天的天气晴朗, 有 80% 的机会明天也是晴天, 不会下雨; 如果是阴天的话, 有 20% 的机会明天会下雨, 而明天会变晴的机会有 30%. 如果今天下雨, 明天也下雨的概率为 0.8, 但也有 10% 的机会明天会变晴.

(a) 求 MiniGolf 球场的期望日收入额. (b) 求天气不是晴天的平均天数.

2. Joe 喜欢去饭馆吃饭, 他最喜欢的菜有墨西哥菜、意大利菜、中国菜和泰国菜. Joe 吃一顿墨西哥餐要付 \$10.00, 吃一顿意大利餐需要 \$15.00, 吃一顿中国菜得 \$9.00, 而吃泰国饭要 \$11.00. Joe 的饮食习惯很有规律: 有 70% 的机会, 今天的饭是重复昨天的口味, 而换成其他 3 种风味餐的概率相同.

(a) 他每天的餐费平均要花多少钱? (b) 隔几天 Joe 就要吃一次墨西哥餐?

3. 某些被判过刑的人按四种状态之一度过他们的余生: 自由、受审、蹲监狱、缓刑. 每年初的统计表明, 有 50% 的机会一个服过刑的人会再犯罪并且受到审判. 法官可能会把他送进监狱的概率为 0.6, 获得缓刑的概率是 0.4. 一旦进了监狱, 有 10% 的犯人因为表现良好而获得释放. 而在缓刑犯人中, 有 10% 的人又犯了新罪而受到新的审判, 其中有 50% 的人因违犯缓刑令被抓回继续服刑, 有 10% 的人因缺少证据而被释放. 由纳税人支付惩处犯人有关的费用. 据估计, 一场刑事审判的费用为 \$5 000, 一次监禁的判刑的平均费用是 \$20 000, 缓刑的平均费用是 \$2 000.

(a) 求每个犯人的期望费用.

(b) 隔多长时间一个犯人会返回监狱? 上法庭受审? 释放?

4. 一家商店出售一种特殊商品, 每日需求量可用下面的 pdf 表示:

日需求量 D	0	1	2	3
$P\{D\}$	0.1	0.3	0.4	0.2

该商店正在对两种订货策略进行比较: (1) 若库存水平小于 2, 则每 3 天订 3 件商品, 否则不订; (2) 如库存为零, 每 3 天订货 3 件, 否则不订. 每次送货的固定订货费为 \$300, 多余商品的存货费是每件每天 \$3. 采用立即送货方式.

(a) 为了让订货和存货的总期望日费用最少, 该商店应采用哪种订货策略?

(b) 对这两种订货策略, 比较连续两次用完库存之间的平均天数.

***5.** 在美国有 3 种所得税纳税人: 从来不逃税的人、偶尔逃税的人和总是逃税的人. 从一年到下一年的税收审计检查发现, 那些从不逃税的人中, 有 95% 的人明年还属于这一类, 有 4% 的人会变到 "偶尔逃税" 类, 而且剩下的那一部分人变成了 "总是逃税" 类. 对那些偶尔逃税的人, 6% 的人变成 "从不逃税" 的人, 有 90% 的人还是老样子, 而 4% 的人则变成了 "总是逃税". 而对于总是逃税的那些人, 对应上面的百分比分别为 0%, 10%, 90%.

(a) 用马尔可夫链表示这个问题.

(b) 经过多次转移, "从不逃税"、"偶尔逃税" 和 "总是逃税" 这 3 类人员所占的百分比分别是多少?

(c) 统计表明, "偶尔逃税" 类的纳税人每次逃税约 \$5 000, "总是逃税" 类的人群逃税额约为 \$12 000. 假定平均收入税税率为 12%, 纳税人口有 7 000 万人, 求由于逃税造成的年税收损失.

6. Warehouzer 拥有一块可更替的林地用来种松树. 树木按照树龄分成四类: 树苗 (0~5 年)、幼树 (5~10 年)、成树 (11~15 年) 和老树 (15 年以上). 10% 的树苗和幼树会在到达下一个树龄组之前死亡. 对于成树和老树, 有 50% 成材, 只有 5% 死亡. 由于经营方式的更替性质, 所有成材的和死亡的树木在 5 年周期末都要用新树 (苗) 替换.

(a) 用马尔可夫链表示这个树林更替问题.

(b) 若该林地总共可种 500 000 棵树, 求树林长期的树木类型构成.

(c) 如果每棵树的植树成本为 $1, 一棵成树的市场价值是 $20, 求树林经营的平均年收入.

7. 人口的动态变化受到人们为了寻求更好的生活质量或就业机会而不断迁移的影响. Mobile 市有市内人口、郊区人口和附近的农村人口. 每 10 年进行一次的人口普查结果表明, 有 10% 的农村人口搬到郊区住, 5% 的人搬到市内. 对于郊区人口, 有 30% 的人搬到农村地区, 而且有 15% 的人搬到市内. 市内人口不会搬到郊区, 但有 20% 的人会搬到农村去过平静的生活.

(a) 用马尔可夫链表示这一人口动态问题.

(b) 若整个 Mobile 地区现有 20 000 农村居民、100 000 郊区人口、30 000 市内居民, 那么 10 年以后、20 年以后人口的分布情况会如何?

(c) 求 Mobile 市长期的人口状况.

8. 一家租车公司在凤凰城、丹佛、芝加哥和亚特兰大设有分公司. 该公司允许单程和往返租车方式, 使得在一地租的车可以异地还车. 据统计, 每个周末有 70% 的租车是往返的, 而对于单程租车的情况是: 从凤凰城租的车有 20% 是去丹佛的, 有 60% 是去芝加哥的, 其他的去亚特兰大; 从丹佛租的车中, 有 40% 是去亚特兰大的, 有 60% 去芝加哥; 从芝加哥租的车中, 有 50% 是去亚特兰大的, 其他的是去丹佛的; 而从亚特兰大租的车中, 80% 去了芝加哥, 10% 去丹佛, 10% 开往凤凰城.

(a) 将这一情形表达为马尔可夫链.

(b) 假如该公司每周开始在每个地点投放 100 辆车的话, 两周后车子的分布情况如何?

(c) 若每个地点指定最多接受 110 辆车, 任何一个分公司会出现经常性的位置不够的问题吗?

(d) 求一辆车返回到原先地点需要的平均周数.

9. 一家书店每天开始时都要检查库存情况, 补充到 100 本的水平. 过去 30 天的数据提供了下面的日末库存量: 1, 2, 0, 3, 2, 1, 0, 0, 3, 0, 1, 1, 3, 2, 3, 3, 2, 1, 0, 2, 0, 1, 3, 0, 0, 3, 2, 1, 2, 2.

(a) 将每天的库存表示为一个马尔可夫链.

(b) 求该书店任何一天无书可卖的稳态概率.

(c) 求期望日库存量.

(d) 求连续两次零库存之间的平均天数.

10. 在第 9 题中, 假设日需求量可超出供应量, 从而引起短缺 (负库存). 过去 30 天每天末的库存水平如下: 1, 2, 0, −2, 2, 2, −1, −1, 3, 0, 0, 1, −1, −2, 3, 3, −2, −1, 0, 2, 0, −1, 3, 0, 0, 3, −1, 1, 2, −2.

(a) 将这一情形表示为一个马尔可夫链.

(b) 求任何一天库存出现过剩的长期概率.

(c) 求任何一天库存出现短缺的长期概率.

(d) 求任何一天日供应量刚好满足日需求量的长期概率.

(e) 若每天末每本剩余书的存储成本为 $1.5, 每天每本短缺书的惩罚费为 $4.00, 求每天的期望库存费用.

11. 一家商店在一周开始时有至少 3 台电脑, 每周需求量估计为 0 台的概率是 0.15, 1 台的概率是 0.2, 2 台的概率为 0.35, 3 台的概率是 0.25, 4 台的概率为 0.05. 需求不能满足时可以倒冲. 该商店的订货策略是, 一旦库存水平下降到 3 台电脑, 则要求下周一时送货, 新的补充总会使库存恢复到 5 台电脑的水平上.

(a) 将这一情形表示为一个马尔可夫链.

(b) 假定这周开始有 4 台电脑, 求两周末订货的概率.

(c) 求任何一周没有订货的长期概率.

(d) 假设一次订货的固定订货费是 \$200, 每台电脑每周的存储费为 \$5, 并且每周每台缺货电脑的惩罚费用是 \$20, 求每周的期望库存费用.

12. 假定每次订货时的订货量刚好是 5 台, 求解第 11 题.

13. 在第 12 题中, 假设对 0~5 台电脑的需求量是等概率的, 进一步假设未满足的需求不倒冲, 但仍然发生惩罚费用.

(a) 将这一情形表示为一个马尔可夫链. (b) 求发生缺货的长期概率.

(c) 假设一次订货的固定订货费是 \$200, 每台电脑每周的存储费为 \$5, 并且每周每台缺货电脑的惩罚费用是 \$20, 求每周的期望订货费用和库存费用.

***14.** 联邦政府希望通过发放年度项目经费来促进小企业发展, 项目采用竞争招标方式, 但如果企业在过去的 3 年期间没有得到过任何经费, 则获得经费的机会最高, 而如果企业过去 3 年中每年都得到过经费, 则机会最低. 具体来说, 过去 3 年从未获得过经费情况下能得到资助的概率为 0.9, 若得到过一次经费, 这一概率下降到 0.8, 得到过 2 次经费的下降为 0.7, 而得到过 3 次资助的, 申请到本次经费的概率只有 0.5.

(a) 将这一情形表示为一个马尔可夫链.

(b) 求每家企业每年获得经费的期望次数.

15. Jim Bob 过去吃过许多驾车违章罚款. 不幸的是, 当今的技术能跟踪到这些过去的罚款纪录. 只要累计得到 4 次罚单, 他的驾照就会被吊销, 直到他重新上完一次驾驶员培训班, 他的违章分数才会被清零. 上完驾驶员培训班后, Jim Bob 马上又不在乎了, 有 50% 的机会他又会被警察拦下罚款了. 每次收到新的罚单后, 他才会更加小心一些, 把受罚款的概率降低了 0.1.

(a) 用马尔可夫链表示 Jim Bob 的问题.

(b) 在他的驾照被再次吊销之前, Jim Bob 被警察拦下的平均次数是多少?

(c) Jim Bob 失去驾照的概率是多少?

(d) 假如每次罚款额为 \$100, 被连续两次吊销驾照之间 Jim Bob 要平均被罚多少钱?

17.5 首次通过时间

17.4 节用稳态概率计算了状态 j 的平均首次返回时间 μ_{jj}. 在本节里, 我们讨论如何求**平均首次通过时间** (mean first passage time) μ_{ij}, 即第一次从状态 i 转移到状态 j 所需要的平均转移数. 计算的依据是求从状态 i 到状态 j 至少一次通过的概率 f_{ij}, 计算公式为 $f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)}$, 其中 $f_{ij}^{(n)}$ 是经过 n 次转移从状态 i 到状态

j 第一次通过的概率. $f_{ij}^{(n)}$ 的表达式可从下式递归求出：

$$p_{ij}^{(n)} = f_{ij}^{(n)} + \sum_{k=1}^{n-1} f_{ij}^{(k)} p_{ij}^{(n-k)},\ n = 1, 2, \cdots$$

假定转移矩阵 $\boldsymbol{P} = \|p_{ij}\|$ 有 m 个状态.

(1) 若 $f_{ij} < 1$, 不能确定系统会从状态 i 通过到状态 j, 并且 $\mu_{ij} = \infty$.

(2) 若 $f_{ij} = 1$, 则马尔可夫链是遍历的, 并且从状态 i 到状态 j 的平均首次通过时间为

$$\mu_{ij} = \sum_{n=1}^{\infty} n f_{ij}^{(n)}$$

在 m- 转移矩阵 $\boldsymbol{P}$ 中, 求所有状态的平均首次通过时间的一种更加简单的方法是用下面的矩阵公式：

$$\|\mu_{ij}\| = (\boldsymbol{I} - \boldsymbol{N}_j)^{-1}\boldsymbol{1},\quad j \neq i$$

其中

$\boldsymbol{I} = (m-1)$维单位矩阵

$\boldsymbol{N}_j =$ 转移矩阵 $\boldsymbol{P}$ 去掉目标状态 j 的第 j 行和第 j 列

$\boldsymbol{1} =$ 所有元素为 1 的 $(m-1)$ 维列向量

矩阵运算 $(\boldsymbol{I} - \boldsymbol{N}_j)^{-1}\boldsymbol{1}$ 是将 $(\boldsymbol{I} - \boldsymbol{N}_j)^{-1}$ 的各列相加.

例 17.5-1

再考虑施肥情况下的园艺师问题的马尔可夫链.

$$\boldsymbol{P} = \begin{pmatrix} 0.30 & 0.60 & 0.10 \\ 0.10 & 0.60 & 0.30 \\ 0.05 & 0.40 & 0.55 \end{pmatrix}$$

为了说明从所有其他状态到某个指定状态的首次通过时间的计算过程, 考虑从状态 2 和状态 3(一般和较差) 通过到状态 1(良好). 因此 $j = 1$, 并且

$$\boldsymbol{N}_1 = \begin{pmatrix} 0.60 & 0.30 \\ 0.40 & 0.55 \end{pmatrix},\quad (\boldsymbol{I} - \boldsymbol{N}_1)^{-1} = \begin{pmatrix} 0.4 & -0.3 \\ -0.4 & 0.45 \end{pmatrix}^{-1} = \begin{pmatrix} 7.50 & 5.00 \\ 6.67 & 6.67 \end{pmatrix}$$

因此,

$$\begin{pmatrix} \mu_{21} \\ \mu_{31} \end{pmatrix} = \begin{pmatrix} 7.50 & 5.00 \\ 6.67 & 6.67 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 12.50 \\ 13.34 \end{pmatrix}$$

这说明平均要花 12.5 个季节从一般土壤条件过渡到良好土壤, 并需要 13.34 个季节从较差的土壤变到良好的土壤.

可以进行类似的计算从 $(I-N_2)$ 得到 μ_{12} 和 μ_{32}, 以及从 $(I-N_3)$ 得到 μ_{13} 和 μ_{23}, 如后面所示.

Excel 程序

可以用 Excel 程序 excelFirstPassTime.xls 方便地计算平均首次通过时间. 图 17.2 显示了关于例 17.5-1 的计算结果. 电子表格的 `Step 2` 自动按照 `Step 1` 给定的维数将转移矩阵 P 赋零值. 在 `Step 2a` 中, 可以用自己的编码覆盖第 6 行的默认状态编码, 这些编码然后会自动变换到整个电子表格中. 输入了转移概率以后, `Step 3` 会创建矩阵 $I-P$, `Step 4` 根据 $I-P$ 来创建 $I-N_j$ $(j=1,2,3)$. 可以复制 $I-P$ 和状态代码, 并粘贴到目标位置上, 然后用适当的剪切和粘贴操作去掉 $I-P$ 的第 j 行和第 j 列. 例如, 要创建 $I-N_2$, 首先复制 $I-P$ 和它们的状态编码到选定的目标位置, 接下来, 把已复制的矩阵的第 3 列拉黑, 剪切下来, 并粘贴到第 2 列, 这样就去掉了第 2 列. 类似可把得到的矩阵的第 3 行拉黑, 剪切并粘贴到第 2 行上, 这样就去掉了第 2 行. 所创建的 $(I-N)$ 自动代入了正确的状态编码.

	A	B	C	D	E	F	G	H
1		First Passage Times in Ergodic and Absorbing Markov Chains						
2	Step 1:	Number of states =		3	Step 2a:	You may override codes in ROW 6		
4	Step 2:				Step 3:			
5		Matrix P:(Blank cell may result in "Type mismatch" compiler error)						
6		1	2	3				
7	1	0.3	0.6	0.1				
8	2	0.1	0.6	0.3				
9	3	0.05	0.4	0.55				
10		Matrix I-P:						
11		1	2	3				
12	1	0.7	-0.6	-0.1				
13	2	-0.1	0.4	-0.3				
14	3	-0.05	-0.4	0.45				
15	Step 4: Perform first passage time calculations below:							
16		I-N			inv(I-N)			Mu
17	i=1	2	3		2	3		1
18	2	0.4	-0.3	2	7.5	5	2	12.5
19	3	-0.4	0.45	3	6.666667	6.6666667	3	13.33333
20								
21	i=2	1	3		1	3		2
22	1	0.7	-0.1	1	1.451613	0.3225806		1.774194
23	3	-0.05	0.45	3	0.16129	2.2580645		2.419355
24								
25	i=3	1	2		1	2		3
26	1	0.7	-0.6	1	1.818182	2.7272727		4.545455
27	2	-0.1	0.4	2	0.454545	3.1818182		3.636364

图 17.2 计算例 17.5-1 首次通过时间的 Excel 电子表格程序 (文件 excelFirstPassTime.xls)

一旦创建了 $I-N_j$, 其逆矩阵 $(I-N_j)^{-1}$ 在目标位置上给出. 相应计算如图 17.2 的求 $(I-N_1)$ 的逆所示:

(1) 在 E18 中输入公式 `=MINVERSE(B18:C19)`

(2) 拉黑 E18:F19, 用来放逆矩阵.

(3) 按 F2 键.

(4) 按 CTRL+SHIFT+ENTER.

从状态 2 和状态 3 到状态 1 的首次通过时间的值通过对逆矩阵行的求和来计算，即在 H18 中输入 =SUM(E18:F18)，然后把 H18 复制到 H19. 在对 $i = 2$ 和 $i = 3$ 创建了 $\boldsymbol{I} - \boldsymbol{N}$ 以后，通过复制 E18:F19 到 E22:F23 和 E26:F27，以及复制 H18:H19 到 H22:H23 和 H26:H27 自动进行其他的计算.

习题 17.5A

*1. 一个老鼠迷宫由图 17.3 所示的路径组成. 路口 1 是迷宫的入口，路口 5 是出口. 在任何一个路口上，老鼠等概率地选择已有的路径. 当老鼠到达路口 5 时，允许老鼠在迷宫中再往回走.

(a) 用马尔可夫链表示这个迷宫问题.

(b) 求老鼠从路口 1 开始，经过 3 次试走后到达出口的概率.

(c) 求该老鼠找到出口的长期概率.

(d) 求从路口 1 到达出口点所需要的平均试走次数.

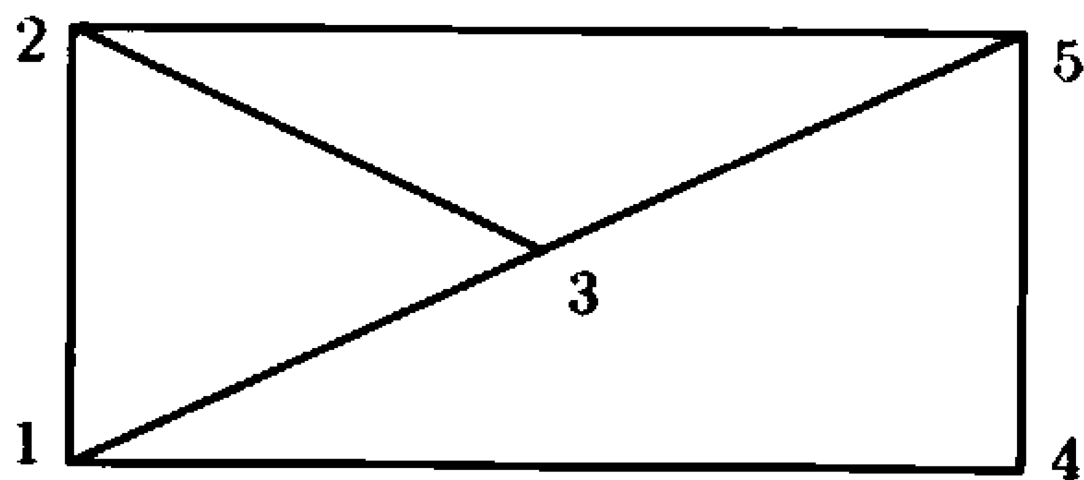


图 17.3 习题 17.5A 第 1 题的老鼠迷宫

2. 在第 1 题中，假如更多的选择（路线）加到迷宫中，到达出口点所需要的平均试走次数会增加还是减少？通过在路口 3 和路口 4 之间增加一条路径来说明你的答案.

3. Jim 和 Joe 用 5 个筹码玩一个游戏，Jim 有 3 个筹码，Joe 有 2 个筹码. 游戏从投掷一枚硬币开始. 若结果是正面朝上，Jim 给 Joe 1 个筹码；否则，Jim 从 Joe 那里得到 1 个筹码. 当 Jim 或 Joe 得到所有的筹码，游戏结束. 现在，有 30% 的机会 Jim 和 Joe 会继续玩这场游戏，还从 Jim 有 3 个筹码、Joe 有 2 个筹码开始.

(a) 用马尔可夫链表示这个游戏.

(b) 求 3 次投掷硬币后 Joe 获胜的概率. 求 3 次投掷硬币后 Jim 获胜的概率.

(c) 求一次游戏结束 Jim 获胜的概率，Joe 获胜的概率.

(d) 求 Jim 获胜之前所需要的平均掷币次数，Joe 获胜之前所需要的平均次数.

4. 一个参加植物学培训的业余园艺师正在进行一项科学实验，把粉色鸢尾花与红色、橙色和白色鸢尾花进行交叉授粉. 他的每年一次的实验表明，粉色可以产生 60% 的粉色和 40% 的白色，红色可以产生出 40% 的红色、50% 的粉色和 10% 的橙色，橙色能产生出 25% 的橙色、50% 的粉色和 25% 的白色，而白色可以产生出 50% 的粉色和 50% 的白色.

(a) 用马尔可夫链表示这个园艺师的实验.

(b) 如果他用同样数目的每种鸢尾花开始交叉授粉实验，5 年以后的分布情况会怎样？长期以后会怎样？

(c) 一株红色的鸢尾花平均需要多少年会开出白花？

*5. 顾客倾向于购买某些特定的产品品牌, 但会受到聪明的推销和广告影响而可能改变对品牌的偏好. 考虑 3 种品牌的情况: A, B, C. 顾客大约 75% 的时候会“坚守”对某个已知品牌的忠诚度, 只给竞争对手 25% 的机会改变品牌. 竞争对手每年开展一次广告宣传活动. 对于品牌 A 的顾客, 改变为购买品牌 B 和品牌 C 的概率分别为 0.1 和 0.15. 品牌 B 的顾客愿意改买品牌 A 和品牌 C 的概率分别有 0.2 和 0.05. 而品牌 C 的顾客等概率地转换到品牌 A 和品牌 B.

(a) 用马尔可夫链表示这一情形.

(b) 长期以后, 每个品牌的市场占有情况如何?

(c) 平均需要多长时间能让品牌 A 的顾客转到品牌 B? 转到到品牌 C?

17.6 对吸收状态的分析

不用施肥的园艺师问题的转移矩阵为

$$
\boldsymbol{P}=\begin{pmatrix} 0.2 & 0.5 & 0.3 \\ 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \end{pmatrix}
$$

状态 1 和状态 2(良好和一般的土壤条件) 是**瞬时的**, 状态 3(较差的土壤条件) 是**吸收的**, 因为一旦处于该状态, 系统将保持不变. 马尔可夫链可以有多个吸收状态. 例如, 一个员工可能一直在同一家公司就职直到退休, 也可能提前几年辞职 (两种吸收状态). 在这类链中, 我们感兴趣的是, 在已知系统从某个指定的瞬时状态出发的情况下, 如何求到达吸收状态的概率和到达吸收状态的期望转移步数. 例如在上面的园艺师问题的马尔可夫链中, 若现在的土壤条件良好, 我们希望知道土壤条件变成较差的平均园艺季节数, 以及相应这一转变的概率是多少.

对带有吸收状态的马尔可夫链可以用矩阵形式进行分析. 首先, 按下面的形式将马尔可夫链进行分块:

$$
\boldsymbol{P}=\left(\begin{array}{c|c} \boldsymbol{N} & \boldsymbol{A} \\ \hline \mathbf{0} & \boldsymbol{I} \end{array}\right)
$$

这个分块矩阵要求所有的吸收状态位于这个新矩阵的右下角. 例如, 考虑下面的转移矩阵:

$$
\boldsymbol{P}=\begin{matrix} \\ 1 \\ 2 \\ 3 \\ 4 \end{matrix}\begin{matrix} \begin{matrix} 1 & \;\;2 & \;\;3 & \;\;4 \end{matrix} \\ \begin{pmatrix} 0.2 & 0.3 & 0.4 & 0.1 \\ 0 & 1 & 0 & 0 \\ 0.5 & 0.3 & 0 & 0.2 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}
$$

这个矩阵 $\boldsymbol{P}$ 可以重新安排并分块为

$$P^* = \begin{array}{c} \\ 1 \\ 3 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{matrix} 1 & 3 & 2 & 4 \end{matrix} \\ \begin{pmatrix} 0.2 & 0.4 & 0.3 & 0.1 \\ 0.5 & 0 & 0.3 & 0.2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

在这种情况下, 我们有

$$N = \begin{pmatrix} 0.2 & 0.4 \\ 0.5 & 0 \end{pmatrix}, \quad A = \begin{pmatrix} 0.3 & 0.1 \\ 0.3 & 0.2 \end{pmatrix}, \quad I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

知道了 A 和 N 的定义以及全 1 元素的单位列向量 $\mathbf{1}$, 可以证明:

$$\text{从状态 } i \text{ 开始到达状态 } j \text{ 中的期望时间} = (I - N)^{-1} \text{的第 } (i,j) \text{ 个元素}$$

$$\text{到吸收状态的期望时间} = (I - N)^{-1}\mathbf{1}$$

$$\text{吸收概率} = (I - N)^{-1}A$$

例 17.6-1[①]

一种产品依次在两台机器 I 和 II 上加工. 在一台机器上加工完成以后, 就进行检查. 有 5% 的机会这件产品在检查之前就发现是废品. 检查后有 3% 的机会发现产品是废品, 有 7% 的机会可以回到同一台机器上返工. 其他情况下, 产品通过这两台机器上的检查.

(a) 一个部件从机器 I 开始加工, 求到每台机器的平均次数.

(b) 假如一批产品有 1 000 件, 从机器 I 开始加工, 可生产出多少合格产品.

对于马尔可夫链, 这个生产过程有 6 种状态: 从机器 I 开始加工 (s1), 机器 I 加工后进行检查 (i1), 开始在机器 II 上加工 (s2), 机器 II 加工后进行检查 (i2), 在 I 或 II 检查后发现是废品 (J), II 检查后发现是合格品 (G). 产品进入 J 和 G 状态就终止了, 因此 J 和 G 是吸收状态. 转移矩阵为:

$$P = \begin{array}{c} \\ s1 \\ i1 \\ s2 \\ i2 \\ J \\ G \end{array} \begin{array}{c} \begin{matrix} s1 & i1 & s2 & i2 & J & G \end{matrix} \\ \left(\begin{array}{cccc|cc} 0 & 0.95 & 0 & 0 & 0.05 & 0 \\ 0.07 & 0 & 0.9 & 0 & 0.03 & 0 \\ 0 & 0 & 0 & 0.95 & 0.05 & 0 \\ 0 & 0 & 0.07 & 0 & 0.03 & 0.9 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \end{array}$$

① 源自 J. Shamblin and G. Stevens, *Operations Research: A Fundamental Approach*, McGraw-Hill, New York, Chapter 4, 1974.

因此有

$$
\boldsymbol{N}=\begin{matrix} \\ \text{s1} \\ \text{i1} \\ \text{s2} \\ \text{i2} \end{matrix}\begin{matrix} \begin{matrix} \text{s1} & \text{i1} & \text{s2} & \text{i2} \end{matrix} \\ \begin{pmatrix} 0 & 0.95 & 0 & 0 \\ 0.07 & 0 & 0.9 & 0 \\ 0 & 0 & 0 & 0.95 \\ 0 & 0 & 0.07 & 0 \end{pmatrix} \end{matrix}, \quad \boldsymbol{A}=\begin{matrix} \begin{matrix} \text{J} & \text{G} \end{matrix} \\ \begin{pmatrix} 0.05 & 0 \\ 0.03 & 0 \\ 0.05 & 0 \\ 0.03 & 0.9 \end{pmatrix} \end{matrix}
$$

用电子表格计算程序 excelEx17.6-1.xls(见例 17.5-1 的*Excel* 程序), 我们得到

$$
(\boldsymbol{I}-\boldsymbol{N})^{-1}=\begin{pmatrix} 1 & -0.95 & 0 & 0 \\ -0.07 & 1 & -0.9 & 0 \\ 0 & 0 & 0 & -0.95 \\ 0 & 0 & -0.07 & 1 \end{pmatrix}^{-1}=\begin{pmatrix} 1.07 & 1.02 & 0.98 & 0.93 \\ 0.07 & 1.07 & 1.03 & 0.98 \\ 0 & 0 & 1.07 & 1.02 \\ 0 & 0 & 0.07 & 1.07 \end{pmatrix}
$$

$$
(\boldsymbol{I}-\boldsymbol{N})^{-1}\boldsymbol{A}=\begin{pmatrix} 1.07 & 1.02 & 0.98 & 0.93 \\ 0.07 & 1.07 & 1.03 & 0.98 \\ 0 & 0 & 1.07 & 1.02 \\ 0 & 0 & 0.07 & 1.07 \end{pmatrix}\begin{pmatrix} 0.05 & 0 \\ 0.03 & 0 \\ 0.05 & 0 \\ 0.03 & 0.9 \end{pmatrix}=\begin{pmatrix} 0.16 & 0.84 \\ 0.12 & 0.88 \\ 0.08 & 0.92 \\ 0.04 & 0.96 \end{pmatrix}
$$

$(\boldsymbol{I}-\boldsymbol{N})^{-1}$ 的头一行给出了一个零部件从机器 I 开始加工到达每个状态的平均次数. 具体来说, 机器 I 为 1.07 次, 检查 I 为 1.02 次, 机器 II 为 0.98 次, 检查 II 是 0.93 次. 到达机器 I 和检查 I 的次数大于 1 的原因是因为返工和重新检查的缘故. 另一方面, 到达机器 II 的次数值小于 1 的原因是, 有些部件在到达机器 II 之前就发现是废品. 的确, 在 (没有发现废品并且不需要返工的) 最佳条件下, 矩阵 $(\boldsymbol{I}-\boldsymbol{N})^{-1}$ 将表明每个状态只到达一次 (通过把所有状态的转移概率都设成 1, 检验这一结论). 当然, 每个状态的停留时间不同. 例如, 假如机器 I 和机器 II 的加工时间分别是 20 分钟和 30 分钟, 并且机器 I 和 II 的检查时间分别是 5 分钟和 7 分钟的话, 则从机器 I 开始的加工时间 (要么是废品, 要么加工完成) 是 $1.07\times20+1.02\times5+0.98\times30+0.93\times7=62.41$ 分钟.

为了求出 1 000 件产品的成品数, 我们可以从 $(\boldsymbol{I}-\boldsymbol{N})^{-1}\boldsymbol{A}$ 的第一行看出:

$$
\text{一件产品是废品的概率}=0.16 \qquad \text{一件产品是成品的概率}=0.84
$$

这意味着我们可以得到 $1\,000\times0.84=840$ 件成品.

习题 17.6A

1. 在例 17.6-1 中, 假设机器 I 和 II 的人工费用为每小时 \$20, 而检查的每小时人工费用只有 \$18. 进一步假定在机器 I 和 II 上加工一件产品分别需要 30 分钟和 20 分钟, 每台机器上的检查时间都是 10 分钟. 求生产一件成品 (合格品) 所需要的总人工费用.

*2. 每当我从市图书馆借出一本书时, 我通常都尽量一周后还回去. 根据书的薄厚和我闲暇时间的多少, 有 30% 的机会我可能会再看一个星期. 假如这本书借期有两个礼拜了, 则有 10%

的机会我会再看一个星期, 但是我绝对不会三个星期以后才还书.

(a) 用马尔可夫链表达这一情形.

(b) 求还给图书馆之前书在我手上的平均周数.

3. 在 Casino del Rio 赌场里, 一个赌客可以按整元下注, 每注要么以 0.4 的概率赢 \$1, 也可能以 0.6 的概率输 \$1. 开赌时他手上有 3 美元, 若所有的钱都输了, 或者赢了一倍, 赌博结束.

(a) 用马尔可夫链表达这一问题. (b) 求到赌博结束前的平均下注数.

(c) 求赌博以他手上有了 \$6 结束的概率. 求输掉所有 \$3 的概率.

4. Jim 必须要花 5 年的努力完成他在 ABC 大学的博士学位. 可是他非常喜欢学生生活, 并不着急完成学业. 在每个学年, 有 50% 的机会他可能会休学, 只有 50% 的机会去修全时学位课. 在完成了 3 个学年以后, 有 30% 的机会 Jim 可能会“放弃了”, 只得到一个硕士学位, 有 20% 的机会下一年休学, 但会继续修博士学位, 剩下 50% 的机会全天上博士课程.

(a) 用马尔可夫链表达这一情形.

(b) 求 Jim 的学生生涯结束前的期望学年数.

(c) 求 Jim 只以硕士学位结束他的学业旅程的概率.

(d) 假设 Jim 有奖学金支付每年 \$15 000 的生活费 (只能在全时上课的情况下), 在他完成学业前花了多少钱?

5. 一个职工现在 55 岁, 打算 62 岁退休, 但不排除提前退休的可能性. 每个年末, 他都会考虑各种可能性 (是否继续工作). 1 年后退休的概率只有 0.1, 但每过了一年都会增加约 0.01.

(a) 用马尔可夫链表达这一问题.

(b) 求这位职工一直留在公司直到 62 岁退休的概率.

(c) 假设在他 57 岁时, 求他提前退休的概率.

(d) 假设在他 58 岁时, 求他退休之前还剩的期望年数.

6. 在习题 17.1A 的第 3 题中, 求下列各值.

(a) 求一笔贷款要么重新偿还要么作为坏账损失的期望季数.

(b) 求一笔新贷款将作为坏账勾销的概率. 求全部偿还的概率.

(c) 设一笔贷款已经 6 个月了, 求直到确定了贷款状况的季度数.

7. 在男单网球锦标赛中, Andre 和 John 进行争夺冠军的决赛, 五局三胜者赢得冠军. 统计表明, 每一局里 Andre 获胜的概率为 60%.

(a) 用马尔可夫链表示这场比赛.

(b) 平均来讲, 这场比赛要打几局? Andre 获得冠军的概率是多少?

(c) 若现在的分数是 1 比 2, John 领先, Andre 获胜的概率有多少?

(d) 在 (c) 中, 求离比赛结束剩下的平均局数, 并解释其结果.

*8. 某大学的学生们对数学系教的一个学期的初等微积分课程进度太快表示不满. 为了解决这一问题, 数学系现在把初等微积分课分成 4 段, 学生们可以对每一段安排自己的进度, 学完后参加一个测验, 根据测验分数决定进入下一个阶段. 测验每 4 个星期进行一次, 以便让勤奋的学生能在一学期内学完 4 个阶段的课程. 按照这种自行安排进度的课程, 几年以后

发现, 有 20% 的学生不能按时学完第一段课程, 不能按时学完第 2 段到第 4 段课程的学生分别占 22%, 25%, 30%.

(a) 用马尔可夫链表示这个问题.

(b) 按照平均, 一个学生从这个学期开始学第一段课程, 他下学期能选高等微积分吗 (初等微积分是基础)?

(c) 一个学生上学期只学完了第一段课程, 他能在本学期完成所有的初等微积分课吗?

(d) 你会建议这种分段教学方式推广到其他的初等数学课吗? 为什么?

9. 在某大学里, 从助理教授晋升到副教授需要相当于 5 年的资历. 每年进行业绩考核, 对申请人给出评语: 一般、良好、优秀. 获得一般评语就等于察看期, 不能得到晋升的资历. 评语良好相当于获得 1 年的资历, 而且得到优秀评语就加上 2 年的资历. 统计表明, 每年有 10% 的申请人被评为一般, 70% 的人被评为良好, 其他的是优秀.

(a) 用马尔可夫链表达这一问题.

(b) 求一位助理教授获得晋升所需要的平均年数.

***10.** (Pfifer and Carraway, 2000) 一家公司通过直接发送邮件广告争取顾客. 在头一年里, 顾客购买一次产品的概率是 0.5, 第 2 年降低到 0.4, 第 3 年为 0.3, 第 4 年减少到 0.2. 如果连续 4 年都没有购买的话, 就把这名顾客就从邮寄名单上去掉, 购买一次就把计数重新设成零.

(a) 用马尔可夫链表示这一情形.

(b) 求一个新顾客留在邮寄名单上的期望年数.

(c) 若一个顾客 2 年内没有购买, 求他留在邮寄名单上的期望年数.

11. 一台数控机床的设计电压标准在 108~112 伏特. 若电压超出这个范围, 机床就停止工作. 这台机床的电压调制器能检测到 1 伏特的增减变化. 经验表明, 每 15 分钟出现一次电压变化, 并且在允许范围内 (108~112 伏) 电压增加 1 伏特、保持不变或者降低 1 伏特的概率相同.

(a) 用马尔可夫链表示这一情形.

(b) 求由于电压低 (高) 造成机床停止工作的概率.

(c) 如何为该机床设定电压, 使得它的工作时间最长?

12. 考虑习题 17.1A 的第 4 题关于患肾衰竭的病人. 求下列各值.

(a) 病人处于透析治疗的期望年数. (b) 一个开始透析治疗的病人的存活年数.

(c) 一个接受了肾移植后存活一年或以上的病人的寿命情况.

(d) 一个肾移植后存活至少一年的病人在透析或死亡前的期望年数.

(e) 那些肾移植后存活一年或以上的病人的生活质量 (假设透析年数少表示生活质量高).

参考文献

Derman, C., *Finite State Markovian Decision Processes,* Academic Press, New York, 1970.

Cyert, R., H. Davidson, and G. Thompson, "Estimation of the Allowance for Doubtful Accounts by Markov Chains," *Management Science,* Vol. 8, No.4, pp.287-303, 1963.

Pfifer, P., and R. Cassaway, "Modeling Customer Relations with Markov Chains," *Journal of Interactive Marketing,* Vol.14, No.2, pp.43-55, 2000.

Grimmet, G., and D. Stirzaker, *Probability and Random Processes,* 2nd ed., Oxford University Press, Oxford, England, 1992.

Kallenberg, O., *Foundations of Modern Probability,* Springer-Verlag, New York, 1997.

Pliskin, J., and E. Tell, "Using Dialysis Need-Projection Model for Health Planning in Massachusetts," *Interfaces,* Vol.11, No.6, pp.84-99, 1981.

Stewart, W., *Introduction to the Numerical Solution of Markov Chains,* Princeton University Press, Princeton, NJ, 1995.

第 18 章　经典最优化理论

本章导读　经典最优化理论利用微分学来求出无约束的和有约束的函数的最大值点和最小值点 (极值点). 这些方法不一定适用于高效率的数值计算, 但这些理论为大部分非线性规划算法 (见第 19 章) 提供了理论基础. 本章为无约束极值问题建立了必要条件和充分条件, 并对等式约束问题给出了雅可比方法和拉格朗日方法, 对不等式约束问题给出了 Karush-Kuhn-Tucker (KKT) 条件, 该条件为所有非线性规划问题提供了最具统一性的理论.

本章包括 10 个例题、1 个电子表格程序和 23 个节后习题. AMPL/Excel/Solver /TORA 程序放在文件夹 ch18Files 下.

18.1　无约束问题

函数 $f(\boldsymbol{X})$ 的极值点定义为该函数的一个极大值点或者极小值点. 用数学公式表示, 一个点 $\boldsymbol{X}_0=(x_1^0,\cdots,x_j^0,\cdots,x_n^0)$ 是一极大值点, 若

$$f(\boldsymbol{X}_0+\boldsymbol{h})\leqslant f(\boldsymbol{X}_0)$$

对所有的 $\boldsymbol{h}=(h_1,\cdots,h_j,\cdots,h_n)$ 都成立, 其中对所有的 j, $|h_j|$ 充分小. 换句话说, 若 f 在 $\boldsymbol{X}_0$ 的邻域内的每个点上的值都不超过 $f(\boldsymbol{X}_0)$, 则 $\boldsymbol{X}_0$ 是一个极大值点. 类似地, $\boldsymbol{X}_0$ 是一极小值点, 若

$$f(\boldsymbol{X}_0+\boldsymbol{h})\geqslant f(\boldsymbol{X}_0)$$

图 18.1 给出了一元函数 $f(x)$ 在区间 $[a,b]$ 上的极大值点和极小值点. 点 x_1, x_2, x_3, x_4, x_6 都是 $f(x)$ 的极值点, 其中 x_1, x_3, x_6 是极大值点, x_2 和 x_4 为极小值点. 因为

$$f(x_6)=\max\{f(x_1),f(x_3),f(x_6)\}$$

所以 $f(x_6)$ 是**全局** (global) 或**绝对** (absolute) 极大值, $f(x_1)$ 和 $f(x_3)$ 为**局部** (local) 或**相对** (relative) 极大值. 类似地, $f(x_4)$ 是局部极小值, 而 $f(x_2)$ 是全局极小值.

虽然 (图 18.1 中的) x_1 是极大值点, 但它和其他的局部极大值点不同, 因为在 x_1 的邻域里至少有一个点相应的 f 值等于 $f(x_1)$. 在这种情况下, 我们称 x_1 是**弱极大值点** (weak maximum), 而称 x_3 和 x_6 是**强极大值点** (strong maxima). 一般地,

对于前面定义的 $\boldsymbol{h}$, $\boldsymbol{X}_0$ 是一个弱极大值点, 如果 $f(\boldsymbol{X}_0+\boldsymbol{h}) \leqslant f(\boldsymbol{X}_0)$; $\boldsymbol{X}_0$ 是一强极大值点, 若 $f(\boldsymbol{X}_0+\boldsymbol{h}) < f(\boldsymbol{X}_0)$.

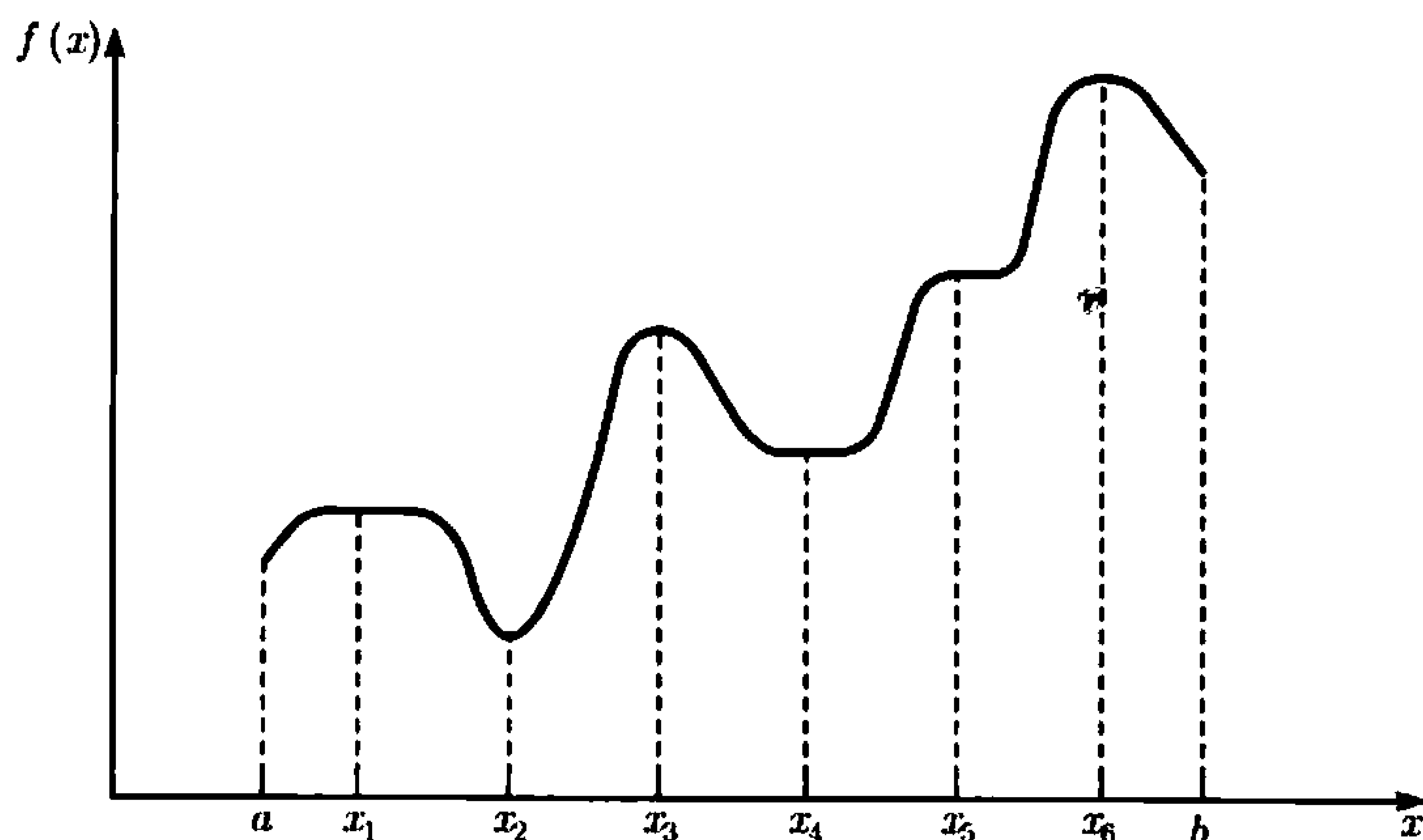


图 18.1　单变量函数极值点的示例

在图 18.1 中, f 在所有的极值点上的一阶导数 (斜率) 等于零, 但是这一性质在**拐点** (inflection) 和**鞍点** (saddle) 处也成立, 正如点 x_5 所示. 若一个零斜率 (梯度) 点不是极值点 (极大值点或极小值点), 则它必是拐点或鞍点.

18.1.1　必要条件和充分条件

本节建立 n 元变量函数 $f(\boldsymbol{X})$ 具有极值点的必要条件和充分条件. 假定, $f(\boldsymbol{X})$ 的一阶和二阶偏导数对所有的 $\boldsymbol{X}$ 都是连续的.

定理 18.1-1　*$\boldsymbol{X}_0$ 为 $f(\boldsymbol{X})$ 的极值点的必要条件是*

$$\nabla f(\boldsymbol{X}_0) = 0$$

证明　根据泰勒 (Taylor) 定理, 对于 $0<\theta<1$,

$$f(\boldsymbol{X}_0+\boldsymbol{h}) - f(\boldsymbol{X}_0) = \nabla f(\boldsymbol{X}_0)\boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{h}|_{\boldsymbol{X}_0+\theta\boldsymbol{h}}$$

对充分小的 $|h_j|$, 余项 $\frac{1}{2}\boldsymbol{h}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{h}$ 与 h_j^2 同阶, 因此

$$f(\boldsymbol{X}_0+\boldsymbol{h}) - f(\boldsymbol{X}_0) = \nabla f(\boldsymbol{X}_0)\boldsymbol{h} + o(h_j^2) \approx \nabla f(\boldsymbol{X}_0)\boldsymbol{h}$$

我们用反证法说明, $\nabla f(\boldsymbol{X}_0)$ 在极小值点 $\boldsymbol{X}_0$ 处必为零. 假设不然, 则对某个 j 下列条件成立:

$$\frac{\partial f(\boldsymbol{X}_0)}{\partial x_j} < 0 \quad \text{或者} \quad \frac{\partial f(\boldsymbol{X}_0)}{\partial x_j} > 0$$

适当选择 h_j 的符号, 总能使得

$$h_j \frac{\partial f(\boldsymbol{X}_0)}{\partial x_j} < 0$$

如果令所有其他的 h_j 为零, 泰勒展开式变成

$$f(\boldsymbol{X}_0+\boldsymbol{h}) < f(\boldsymbol{X}_0)$$

这一结果与 $\boldsymbol{X}_0$ 是极小值点相矛盾. 因此 $\nabla f(\boldsymbol{X}_0)$ 必为零. 对极大值点的情况可得出相类似的证明.

因为在拐点和鞍点处总是满足上述必要条件, 我们把根据 $\nabla f(\boldsymbol{X}_0)=\boldsymbol{0}$ 得到的点称为**稳定** (stationary) 点. 下面的定理建立 $\boldsymbol{X}_0$ 为极值点的充分条件.

定理 18.1-2 *稳定点 $\boldsymbol{X}_0$ 是极值点的充分条件是, 在 $\boldsymbol{X}_0$ 点的 Hesse (黑塞) 矩阵 $\boldsymbol{H}$ 满足下列条件:*

(i) *若 $\boldsymbol{H}$ 是正定的, 则 $\boldsymbol{X}_0$ 是极小值点;*

(ii) *若 $\boldsymbol{H}$ 是负定的, 则 $\boldsymbol{X}_0$ 是极大值点.*

证明 依泰勒定理, 对 $0<\theta<1$,

$$f(\boldsymbol{X}_0+\boldsymbol{h})-f(\boldsymbol{X}_0)=\nabla f(\boldsymbol{X}_0)\boldsymbol{h}+\frac{1}{2}\boldsymbol{h}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{h}|_{\boldsymbol{X}_0+\theta\boldsymbol{h}}$$

由于 $\boldsymbol{X}_0$ 是稳定点, 则 $\nabla f(\boldsymbol{X}_0)=\boldsymbol{0}$ (定理 18.1-1). 因此

$$f(\boldsymbol{X}_0+\boldsymbol{h})-f(\boldsymbol{X}_0)=\frac{1}{2}\boldsymbol{h}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{h}|_{\boldsymbol{X}_0+\theta\boldsymbol{h}}$$

若 $\boldsymbol{X}_0$ 是极小值点, 则

$$f(\boldsymbol{X}_0+\boldsymbol{h}) > f(\boldsymbol{X}_0), \quad \boldsymbol{h}\neq\boldsymbol{0}$$

于是, 欲使 $\boldsymbol{X}_0$ 是极小值点, 则必有

$$\frac{1}{2}\boldsymbol{h}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{h}|_{\boldsymbol{X}_0+\theta\boldsymbol{h}}>0$$

由二阶偏导数的连续性, 对充分小的 $|h_j|$, 表达式 $\frac{1}{2}\boldsymbol{h}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{h}$ 在 $\boldsymbol{X}_0$ 和 $\boldsymbol{X}_0+\theta\boldsymbol{h}$ 点的符号必相同. 由于 $\boldsymbol{h}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{h}|_{\boldsymbol{X}_0}$ 定义一个二次型 (见附录 D.3), 该表达式 (因而 $\boldsymbol{h}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{h}|_{\boldsymbol{X}_0+\theta\boldsymbol{h}}$) 是正的, 当且仅当 $\boldsymbol{H}|_{\boldsymbol{X}_0}$ 是正定的. 这就意味着, 稳定点 $\boldsymbol{X}_0$ 是极小点的充分条件是, 在相同点上的 Hesse 矩阵是正定的. 即由 $\boldsymbol{H}|_{\boldsymbol{X}_0}$ 是正定的, 必有 $\frac{1}{2}\boldsymbol{h}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{h}|_{\boldsymbol{X}_0+\theta\boldsymbol{h}}>0$, 所以 $f(\boldsymbol{X}_0+\boldsymbol{h})>f(\boldsymbol{X}_0)$, $\boldsymbol{h}\neq\boldsymbol{0}$, 则 $\boldsymbol{X}_0$ 是极小点. 对于 Hesse 矩阵为负定的情况, 可证明 $\boldsymbol{X}_0$ 是极大值点.

例 18.1-1

考虑函数

$$f(x_1,x_2,x_3)=x_1+2x_3+x_2x_3-x_1^2-x_2^2-x_3^2$$

必要条件 $\nabla f(\boldsymbol{X}_0)=0$ 给出

$$\frac{\partial f}{\partial x_1}=1-2x_1=0$$

$$\frac{\partial f}{\partial x_2} = x_3 - 2x_2 = 0$$
$$\frac{\partial f}{\partial x_3} = 2 + x_2 - 2x_3 = 0$$

这组方程的解为

$$\boldsymbol{X}_0 = \left(\frac{1}{2}, \frac{2}{3}, \frac{4}{3}\right)$$

为确定该稳定点的类别, 考虑

$$\boldsymbol{H}|_{\boldsymbol{X}_0} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_1 \partial x_3} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \frac{\partial^2 f}{\partial x_2 \partial x_3} \\ \frac{\partial^2 f}{\partial x_3 \partial x_1} & \frac{\partial^2 f}{\partial x_3 \partial x_2} & \frac{\partial^2 f}{\partial x_3^2} \end{pmatrix}_{\boldsymbol{X}_0} = \begin{pmatrix} -2 & 0 & 0 \\ 0 & -2 & 1 \\ 0 & 1 & -2 \end{pmatrix}$$

$\boldsymbol{H}|_{\boldsymbol{X}_0}$ 的主子行列式分别为 $-2, 4, 6$, 因此如附录 D.3 所示, $\boldsymbol{H}|_{\boldsymbol{X}_0}$ 为负定, $\boldsymbol{X}_0 = \left(\frac{1}{2}, \frac{2}{3}, \frac{4}{3}\right)$ 代表极大值点.

一般情况下, 若 $\boldsymbol{H}|_{\boldsymbol{X}_0}$ 为不定的, $\boldsymbol{X}_0$ 必为一鞍点. 对于不确定情况下, $\boldsymbol{X}_0$ 可能是也可能不是极值点, 因此充分条件变得至关重要, 因为这时必需要考虑泰勒展开的更高阶项.

下面给出定理 18.1-2 中的充分条件用在一元变量函数的情形. 已知 y_0 是一个稳定点, 则

(i) 如 $f''(y_0) < 0$, 则 y_0 为极大值点;

(ii) 若 $f''(y_0) > 0$, 则 y_0 为极小值点.

如果 $f''(y_0) = 0$, 则按照下面定理的要求, 必须考察更高阶导数.

定理 18.1-3 *已知 y_0 是 $f(y)$ 的稳定点, 若前 $(n-1)$ 阶导数均为零, 且 $f^{(n)}(y_0) \neq 0$, 则*

(i) 如 n 为奇数, 则 y_0 是拐点;

(ii) 若 n 为偶数, 则当 $f^{(n)}(y_0) > 0$ 时 y_0 是极小值点, 当 $f^{(n)}(y_0) < 0$ 时 y_0 为极大值点.

例 18.1-2

图 18.2 画出了以下两个函数:

$$f(y) = y^4$$
$$g(y) = y^3$$

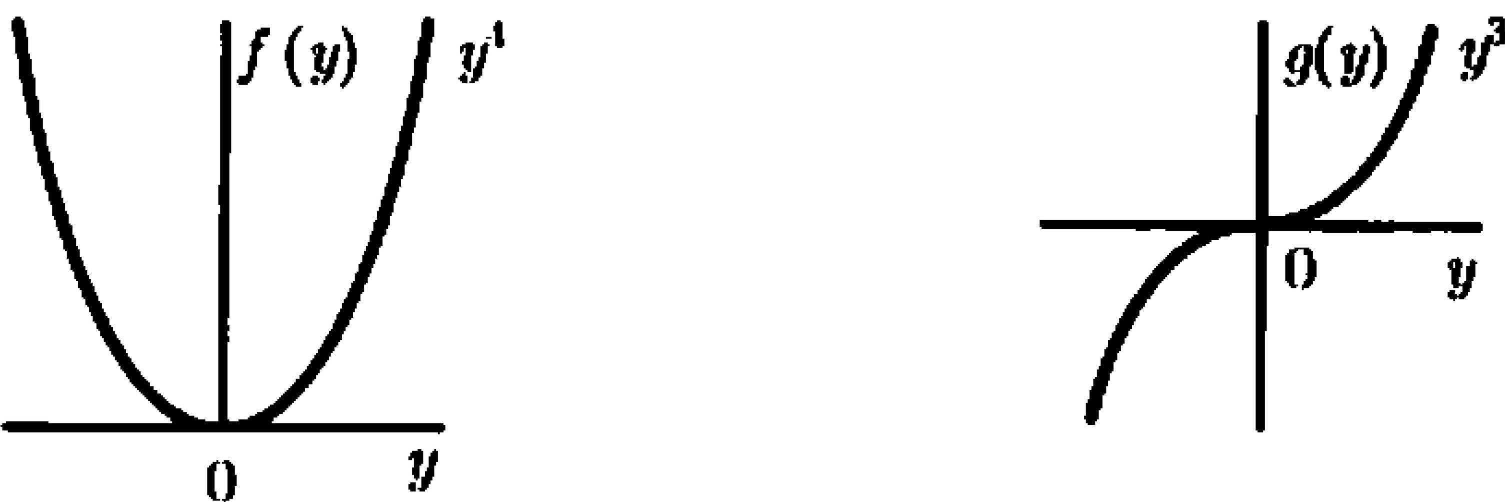


图 18.2 函数 $f(y) = y^4$ 和 $g(y) = y^3$ 的极值点

对 $f(y)=y^4$, $f'(y)=4y^3=0$ 给出稳定点 $y_0=0$. 由于

$$f'(0)=f''(0)=f^{(3)}(0)=0, f^{(4)}(0)=24>0$$

因此, $y_0=0$ 为极小值点 (见图 18.2).

对 $g(y)=y^3, g'(y)=3y^2=0$ 得出 $y_0=0$ 为稳定点, 另外

$$g'(0)=g''(0), g^{(3)}(0)=6\neq 0$$

因此 $y_0=0$ 是一个拐点.

习题 18.1A

1. 求下列函数的极值点.

*(a) $f(x)=x^3+x$　　*(b) $f(x)=x^4+x^2$

(c) $f(x)=4x^4-x^2+5$　　(d) $f(x)=(3x-2)^2(2x-3)^2$

*(e) $f(x)=6x^5-4x^3+10$

2. 求下列函数的极值点.

(a) $f(\boldsymbol{X})=x_1^3+x_2^3-3x_1x_2$

(b) $f(\boldsymbol{X})=2x_1^2+x_2^2+x_3^2+6(x_1+x_2+x_3)+2x_1x_2x_3$

3. 证明函数

$$f(x_1,x_2,x_3)=2x_1x_2x_3-4x_1x_3-2x_2x_3+x_1^2+x_2^2+x_3^2-2x_1-4x_2+4x_3$$

有稳定点 $(0,3,1),(0,1,-1),(1,2,0),(2,1,1),(2,3,-1)$. 利用充分条件找出极值点.

*4. 通过变换成无约束非线性目标函数, 解下列方程组.

$$x_2-x_1^2=0$$

$$x_2-x_1=2$$

[提示: $\min f^2(x_1,x_2)$ 在 $f'(x_1,x_2)=0$ 处达到.]

5. 证明定理 18.1-3.

18.1.2 Newton-Raphson 方法

一般来说, 数值求解必要条件方程 $\nabla f(\boldsymbol{X})=\boldsymbol{0}$ 可能很难. Newton-Raphson 方法是求解非线性方程组的一种迭代法.

考虑下面的方程组:

$$f_i(\boldsymbol{X})=0,\quad i=1,2,\cdots,m$$

令 $\boldsymbol{X}_k$ 为一已知点, 由泰勒展开式:

$$f_i(\boldsymbol{X})\approx f_i(\boldsymbol{X}_k)+\nabla f_i(\boldsymbol{X}_k)(\boldsymbol{X}-\boldsymbol{X}_k),\quad i=1,2,\cdots,m$$

因此, 原来的方程 $f_i(\boldsymbol{X})=0$, $i=1,2,\cdots,m$ 可以近似为

$$f_i(\boldsymbol{X}_k) + \nabla f_i(\boldsymbol{X}_k)(\boldsymbol{X} - \boldsymbol{X}_k) = 0, \quad i = 1, 2, \cdots, m$$

此方程写成以下矩阵形式：

$$\boldsymbol{A}_k + \boldsymbol{B}_k(\boldsymbol{X} - \boldsymbol{X}_k) = \boldsymbol{0}$$

若 $\boldsymbol{B}_k$ 非奇异, 则

$$\boldsymbol{X} = \boldsymbol{X}_k - \boldsymbol{B}_k^{-1}\boldsymbol{A}_k$$

该方法的思路是, 从初始点 $\boldsymbol{X}_0$ 开始, 利用上面的方程组求得一个新点, 继续这一过程直到两个连续的点 $\boldsymbol{X}_k$ 和 $\boldsymbol{X}_{k+1}$ 近似相等.

图 18.3 给出了该方法对一元变量函数的几何解释. 对于一元变量函数 $f(x)$, x_k 和 x_{k+1} 之间的关系变成

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

或

$$f'(x_k) = \frac{f(x_k)}{x_k - x_{k+1}}$$

图像表明, x_{k+1} 由 $f(x)$ 在 x_k 点的斜率求出, 其中 $\tan\theta = f'(x_k)$.

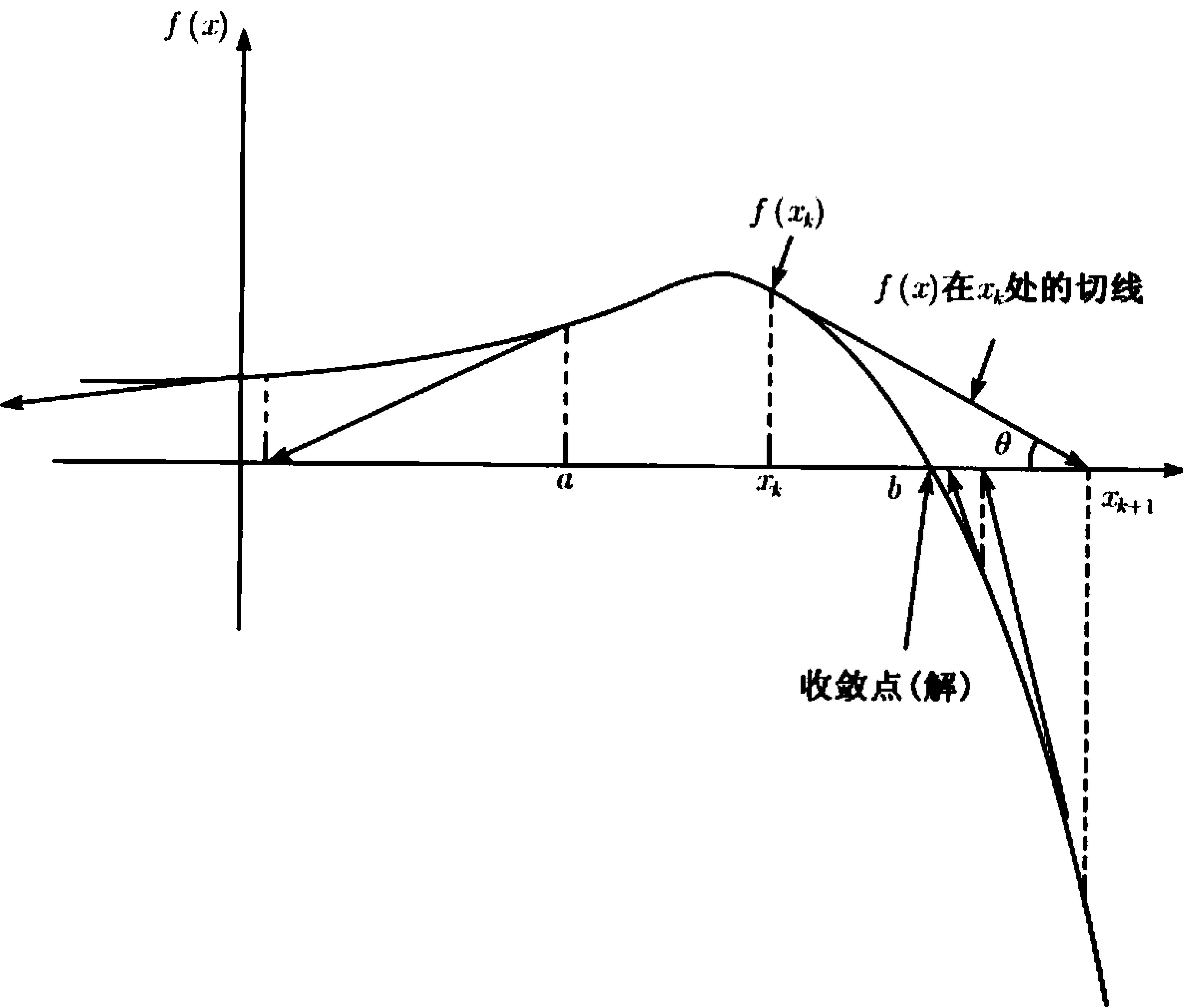


图 18.3　Newton-Raphson 方法的迭代过程示意图

该方法的一个难点是, 当 f 的性质不佳时不能保证收敛. 在图 18.3 中, 若初始点为 a 时, 该方法就会发散. 一般来讲, 需要采用试错的方法, 以便找到一个"好的"初始点.

例 18.1-3

为说明如何使用 Newton-Raphson 方法, 考虑求下列函数的稳定点:

$$g(x) = (3x-2)^2(2x-3)^2$$

要求出稳定点, 需要解方程

$$f(x) \equiv g'(x) = 72x^3 - 234x^2 + 241x - 78 = 0$$

因此对 Newton-Raphson 方法, 我们有

$$f'(x) = 216x^2 - 468x + 241$$
$$x_{k+1} = x_k - \frac{72x^3 - 234x^2 + 241x - 78}{216x^2 - 468x + 241}$$

从 $x_0 = 10$ 开始, 下表提供了连续的迭代结果:

k	x_k	$\frac{f(x_k)}{f'(x_k)}$	x_{k+1}
0	10.000 000	2.978 923	7.032 108
1	7.032 108	1.976 429	5.055 679
2	5.055 679	1.314 367	3.741 312
3	3.741 312	0.871 358	2.869 995
4	2.869 995	0.573 547	2.296 405
5	2.296 405	0.371 252	1.925 154
6	1.925 154	0.230 702	1.694 452
7	1.694 452	0.128 999	1.565 453
8	1.565 453	0.054 156	1.511 296
9	1.511 296	0.010 864 1	1.500 432
10	1.500 432	0.000 431 31	1.500 001

该方法收敛到 $x = 1.5$. 实际上, $f(x)$ 在 $x = \frac{2}{3}, \frac{13}{12}, \frac{3}{2}$ 有 3 个稳定点. 剩下的 2 个点可通过选择不同的初始点 x_0 找出来, 事实上, 从 $x_0 = 0.5$ 和 $x_0 = 1$ 就可以得出另外的 2 个稳定点.

Excel 程序

程序 excelNR.xls 可用来求解任何一元变量方程. 程序要求在单元格 C3 输入 $f(x)/f'(x)$, 如对于例 18.1-3, 我们输入

```
=(72*A3^3-234*A3^2+241*A3-78)/(216*A3^2-468*A3+241)
```

变量 x 用 A3 替代, 程序允许设置一个容许上限值 Δ, 它指定 x_k 和 x_{k+1} 之间的可允许差作为迭代终止的信号. 你最好试着用不同的 x_0 来感受该方法的工作原理.

一般地, 要想得到 "所有的" 解, 需要使用多次 Newton-Raphson 方法. 在例 18.1-3 中, 我们事先知道方程有 3 个根. 而对于复杂的或者多元变量的问题, 情况并非如此.

习题 18.1B

1. 用程序 NewtonRaphson.xls 求解习题 18.1A 的第 1(c) 题.
2. 用 Newton-Raphson 方法求解习题 18.1A 的第 2(b) 题.

18.2 约束问题

本节介绍带有约束的连续函数的最优化问题. 18.2.1 节介绍等式约束的情况, 18.2.2 节解决不等式约束问题. 18.2.1 节介绍的大部分内容可参见 Beightler 等人 (1979, pp. 45-55).

18.2.1 等式约束问题

本节介绍两种方法: **雅可比方法和拉格朗日乘子方法, 拉格朗日方法**可从雅可比方法推出. 这种关系提供了拉格朗日方法的一个有意思的经济学解释.

有约束的导数 (雅可比) 方法 考虑问题

$$\begin{aligned} \min \quad & z = f(\boldsymbol{X}) \\ \text{s.t.} \quad & \boldsymbol{g}(\boldsymbol{X}) = \boldsymbol{0} \end{aligned}$$

其中

$$\boldsymbol{X} = (x_1, x_2, \cdots, x_n), \qquad \boldsymbol{g} = (g_1, g_2, \cdots, g_m)^{\mathrm{T}}$$

函数 $f(\boldsymbol{X})$, $g_i(\boldsymbol{X})$, $i = 1, 2, \cdots, m$ 都是二次连续可微的.

采用有约束的导数的思路是, 在满足约束 $\boldsymbol{g}(\boldsymbol{X}) = \boldsymbol{0}$ 的所有点上, 对 $f(\boldsymbol{X})$ 的一阶偏导数建立一个闭形式的表达式, 使得这些偏导数等于零的点即为对应的稳定点. 然后再用 18.1 节得到的充分条件来检验这些稳定点的极值性.

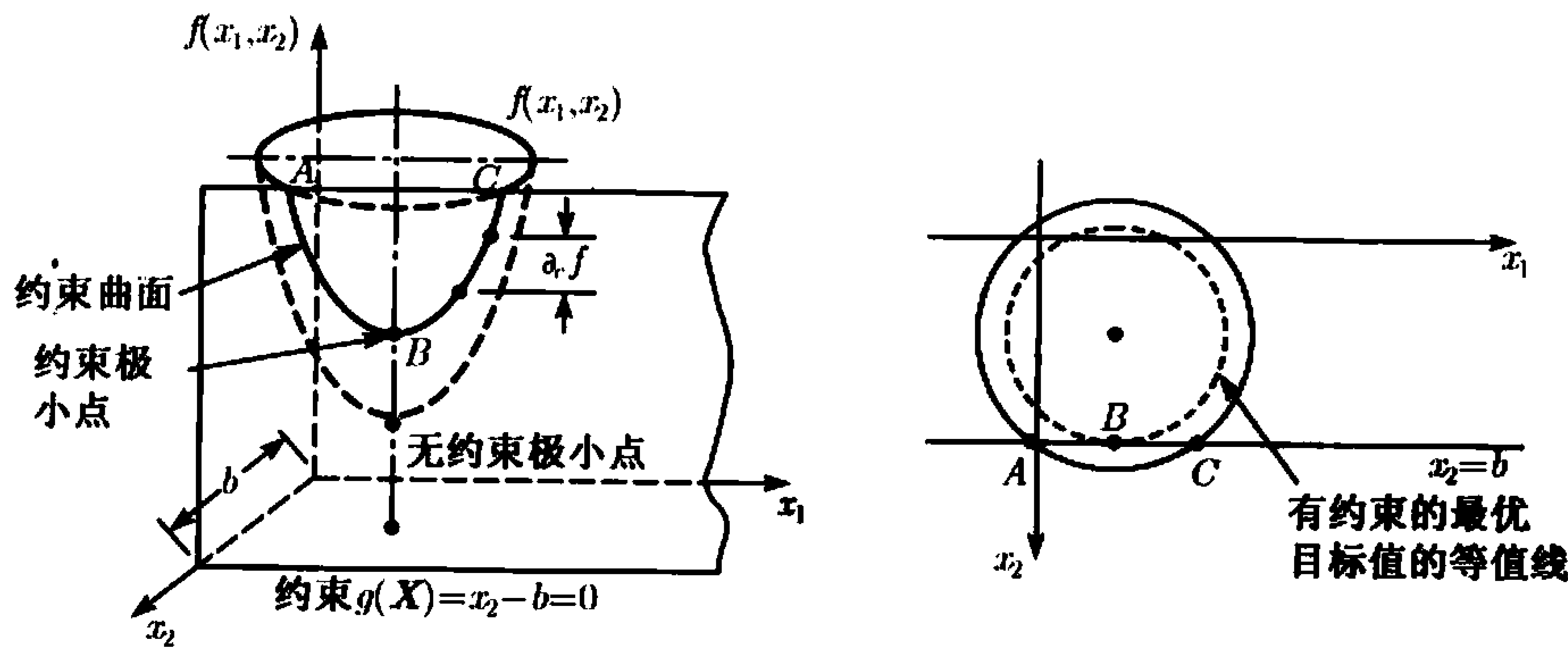


图 18.4 雅可比方法原理的说明

为了进一步说明提出的概念, 考虑图 18.4 展示的函数 $f(x_1, x_2)$, 该函数在约束

$$g_1(x_1, x_2) = x_2 - b = 0$$

下求最小值, 其中 b 为常数. 由图 18.4 可以看出, 由 3 个点 A, B, C 所确定的曲线表示满足给定约束的 $f(x_1, x_2)$ 的值. 有约束的导数方法定义了 $f(x_1, x_2)$ 在曲线 ABC 上任何一点上的梯度, 有约束的导数等于零的点 B 为该约束问题的稳定点.

现在从数学上推导这一方法. 利用泰勒定理, 对 $\boldsymbol{X}$ 可行域内的 $\boldsymbol{X} + \Delta\boldsymbol{X}$, 我们有

$$f(\boldsymbol{X} + \Delta\boldsymbol{X}) - f(\boldsymbol{X}) = \nabla f(\boldsymbol{X})\Delta\boldsymbol{X} + O(\Delta x_j^2)$$

$$\boldsymbol{g}(\boldsymbol{X} + \Delta\boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{X}) = \nabla \boldsymbol{g}(\boldsymbol{X})\Delta\boldsymbol{X} + O(\Delta x_j^2)$$

当 $\Delta x_j \to 0$, 公式变成

$$\partial f(\boldsymbol{X}) = \nabla f(\boldsymbol{X})\partial\boldsymbol{X}, \qquad \partial \boldsymbol{g}(\boldsymbol{X}) = \nabla \boldsymbol{g}(\boldsymbol{X})\partial\boldsymbol{X}$$

为保证可行性, 必须令 $\boldsymbol{g}(\boldsymbol{X}) = \boldsymbol{0}$, $\partial\boldsymbol{g}(\boldsymbol{X}) = \boldsymbol{0}$, 因此有

$$\partial f(\boldsymbol{X}) - \nabla f(\boldsymbol{X})\partial\boldsymbol{X} = \boldsymbol{0}$$

$$\nabla \boldsymbol{g}(\boldsymbol{X})\partial\boldsymbol{X} = \boldsymbol{0}$$

这是 $(n+1)$ 个未知变量 $\partial f(\boldsymbol{X})$ 和 $\partial\boldsymbol{X}$ 的 $(m+1)$ 个方程. 注意到 $\partial f(\boldsymbol{X})$ 是一个相关变量, 因此一旦 $\partial\boldsymbol{X}$ 已知便可确定它. 这意味着我们有 n 个未知变量的 m 个方程.

若 $m > n$, 则至少有 $(m-n)$ 个等式是冗余的. 通过消除冗余, 该方程组可简化为 $m \leqslant n$ 个等式. 如果 $m = n$, 则解为 $\partial\boldsymbol{X} = \boldsymbol{0}$ 且 $\boldsymbol{X}$ 没有可行域, 即解空间只有一点. 对余下的情形, $m < n$, 则需要进一步推导.

定义

$$\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{Z})$$

使得

$$\boldsymbol{Y} = (y_1, y_2, \cdots, y_m), \quad \boldsymbol{Z} = (z_1, z_2, \cdots, z_{n-m})$$

向量 $\boldsymbol{Y}$ 和 $\boldsymbol{Z}$ 分别称为**相关变量**和**无关变量**. 重新写出用 $\boldsymbol{Y}$ 和 $\boldsymbol{Z}$ 表示的 f 和 $\boldsymbol{g}$ 的梯度向量, 我们得到

$$\nabla f(\boldsymbol{Y}, \boldsymbol{Z}) = (\nabla_{\boldsymbol{Y}} f, \nabla_{\boldsymbol{Z}} f)$$

$$\nabla \boldsymbol{g}(\boldsymbol{Y}, \boldsymbol{Z}) = (\nabla_{\boldsymbol{Y}} \boldsymbol{g}, \nabla_{\boldsymbol{Z}} \boldsymbol{g})$$

定义

$$\boldsymbol{J} = \nabla_{\boldsymbol{Y}} \boldsymbol{g} = \begin{pmatrix} \nabla_{\boldsymbol{Y}} g_1 \\ \vdots \\ \nabla_{\boldsymbol{Y}} g_m \end{pmatrix} \qquad \boldsymbol{C} = \nabla_{\boldsymbol{Z}} \boldsymbol{g} = \begin{pmatrix} \nabla_{\boldsymbol{Z}} g_1 \\ \vdots \\ \nabla_{\boldsymbol{Z}} g_m \end{pmatrix}$$

$J_{m\times n}$ 称为**雅可比矩阵**, $C_{m\times n-m}$ 称为**控制矩阵** (control matirx). 因为给定的 m 个方程按定义为独立的, 所以雅可比矩阵 J 非奇异. 向量 Y 的元素从 X 中选择, 并使得 J 是非奇异的.

原来 $\partial f(X)$ 和 ∂X 中的方程组可重新写成

$$\partial f(Y,Z) = \nabla_Y f\partial Y + \nabla_Z f\partial Z$$

且

$$J\partial Y = -C\partial Z$$

由于 J 非奇异, 所以存在逆矩阵 J^{-1}, 因此,

$$\partial Y = -J^{-1}C\partial Z$$

$\partial f(X)$ 的方程中用 ∂Y 替代得到 ∂f 作为 ∂Z 的函数, 即

$$\partial f(Y,Z) = (\nabla_Z f - \nabla_Y f J^{-1}C)\partial Z$$

从这个公式, 得到关于独立向量 Z 的约束导数:

$$\nabla_c f = \frac{\partial_c f(Y,Z)}{\partial_c Z} = \nabla_Z f - \nabla_Y f J^{-1}C$$

其中 $\nabla_c f$ 称为 f 关于 Z 的**约束梯度**(constrained gradient) 向量, 因此 $\nabla_c f(Y,Z)$ 在稳定点处必为零.

这一充分条件与 18.1 节的类似. Hesse 矩阵对应于独立向量 Z, 并且 Hesse 矩阵的元素必须为约束二阶导数. 为了说明如何得到这一条件, 令

$$\nabla_c f = \nabla_Z f - WC$$

这样就得到, (约束的)Hesse 矩阵的第 i 行为 $\frac{\partial \nabla_c f}{\partial z_i}$. 注意到 W 是 Y 的函数, Y 又是 Z 的函数, 因此根据下面的链式规则可求出 $\nabla_c f$ 关于 z_i 的偏导数:

$$\frac{\partial w_j}{\partial z_i} = \frac{\partial w_j}{\partial y_j}\frac{\partial y_j}{\partial z_i}$$

例 18.2-1

考虑下面的问题:

$$f(X) = x_1^2 + 3x_2^2 + 5x_1x_3^2$$
$$g_1(X) = x_1x_3 + 2x_2 + x_2^2 - 11 = 0$$
$$g_2(X) = x_1^2 + 2x_1x_2 + x_3^2 - 14 = 0$$

给定可行点 $\boldsymbol{X}^0=(1,2,3)$, 我们考察在 $\boldsymbol{X}^0$ 可行域内 $f(=\partial_c f)$ 的变化情况.

令

$$\boldsymbol{Y}=(x_1,x_3),\quad \boldsymbol{Z}=x_2$$

因此有

$$\nabla_{\boldsymbol{Y}} f=\left(\frac{\partial f}{\partial x_1},\frac{\partial f}{\partial x_3}\right)=\left(2x_1+5x_3^2,10x_1x_3\right)$$

$$\nabla_{\boldsymbol{Z}} f=\frac{\partial f}{\partial x_2}=6x_2$$

$$\boldsymbol{J}=\begin{pmatrix}\frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_3}\\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_3}\end{pmatrix}=\begin{pmatrix}x_3 & x_1\\ 2x_1+2x_2 & 2x_3\end{pmatrix}$$

$$\boldsymbol{C}=\begin{pmatrix}\frac{\partial g_1}{\partial x_2}\\ \frac{\partial g_2}{\partial x_2}\end{pmatrix}=\begin{pmatrix}2x_2+2\\ 2x_1\end{pmatrix}$$

假设我们需要在可行点 $\boldsymbol{X}_0=(1,2,3)$ 的可行域内对独立变量 x_2 的微小变化 $\partial x_2=0.01$ 来估计 $\partial_c f$. 我们有

$$\boldsymbol{J}^{-1}\boldsymbol{C}=\begin{pmatrix}3 & 1\\ 6 & 6\end{pmatrix}^{-1}\begin{pmatrix}6\\ 2\end{pmatrix}=\begin{pmatrix}\frac{6}{12} & -\frac{1}{12}\\ -\frac{6}{12} & \frac{3}{12}\end{pmatrix}\begin{pmatrix}6\\ 2\end{pmatrix}\approx\begin{pmatrix}2.83\\ -2.50\end{pmatrix}$$

因此约束 f 的增加值为

$$\partial_c f=\left(\nabla_{\boldsymbol{Z}} f-\nabla_{\boldsymbol{Y}} f\boldsymbol{J}^{-1}\boldsymbol{C}\right)\partial\boldsymbol{Z}=\left(6(2)-(47.30)\begin{pmatrix}2.83\\ -2.50\end{pmatrix}\right)\partial x_2=-46.01\partial x_2$$

给定无关变量 x_2 的 ∂x_2 值, 相关变量 x_1 和 x_3 的可行的 ∂x_1, ∂x_3 值可由下面的公式给出:

$$\partial\boldsymbol{Y}=-\boldsymbol{J}^{-1}\boldsymbol{C}\partial\boldsymbol{Z}$$

因此对 $\partial x_2=0.01$,

$$\begin{pmatrix}\partial x_1\\ \partial x_3\end{pmatrix}=-\boldsymbol{J}^{-1}\boldsymbol{C}\partial x_2=\begin{pmatrix}-0.028\,3\\ 0.025\,0\end{pmatrix}$$

对于给定的 $\partial x_2=0.01$, 我们现在将计算得到的 $\partial_c f$ 值与差值 $f(\boldsymbol{X}_0+\partial\boldsymbol{X})-f(\boldsymbol{X})$ 进行比较.

$$\boldsymbol{X}_0+\partial\boldsymbol{X}=(1-0.028\,3,2+0.01,3+0.025)=(0.971\,7,2.01,3.025)$$

由此得到

$$f(\boldsymbol{X}_0)=58,\quad f(\boldsymbol{X}_0+\partial\boldsymbol{X})=57.523$$

即

$$f(\boldsymbol{X}_0+\partial\boldsymbol{X})-f(\boldsymbol{X}_0)=-0.477$$

量 -0.477 和 $\partial_c f=-46.01\partial x_2=-0.4601$ 非常接近, 这两个值之间的差是由在 $\boldsymbol{X}_0$ 处计算 $\partial_c f$ 的线性近似所引起的.

习题 18.2A

1. 考虑例 18.2-1.
   (a) 用例中给出的两种方法计算 $\partial_c f$, 并用 $\partial x_2=0.001$ 代替 $\partial x_2=0.01$ 计算. 随着 ∂x_2 减小, 线性近似的效果会变得更加不明显了吗?
   *(b) 说明在可行点 $\boldsymbol{X}_0=(1,2,3)$ 处保证点 $\boldsymbol{X}_0+\partial\boldsymbol{X}$ 可行的 $\partial\boldsymbol{X}=(\partial x_1,\partial x_2,\partial x_3)$ 的元素之间的关系.
   (c) 若 $\boldsymbol{Y}=(x_2,x_3)$, $\boldsymbol{Z}=x_1$, 那么 ∂x_1 应是多少, 才会产生例中给出的相同 $\partial_c f$ 值?

例 18.2-2

这个例子说明如何使用约束导数. 考虑问题

$$\begin{aligned}\min\ & f(\boldsymbol{X})=x_1^2+x_2^2+x_3^2\\ \text{s.t.}\ & g_1(\boldsymbol{X})=x_1+x_2+3x_3-2=0\\ & g_2(\boldsymbol{X})=5x_1+2x_2+x_3-5=0\end{aligned}$$

如下求约束极值点. 令

$$\boldsymbol{Y}=(x_1,x_2),\quad \boldsymbol{Z}=x_3$$

以此计算

$$\nabla_{\boldsymbol{Y}}f=\left(\frac{\partial f}{\partial x_1},\frac{\partial f}{\partial x_2}\right)=(2x_1,2x_2),\quad \nabla_{\boldsymbol{Z}}f=\frac{\partial f}{\partial x_3}=2x_3$$

$$\boldsymbol{J}=\begin{pmatrix}1&1\\5&2\end{pmatrix},\quad \boldsymbol{J}^{-1}=\begin{pmatrix}-\frac{2}{3}&\frac{1}{3}\\\frac{5}{3}&-\frac{1}{3}\end{pmatrix},\quad \boldsymbol{C}=\begin{pmatrix}3\\1\end{pmatrix}$$

因此有

$$\nabla_c f=\frac{\partial_c f}{\partial_c x_3}=2x_3-(2x_1,2x_2)\begin{pmatrix}-\frac{2}{3}&\frac{1}{3}\\\frac{5}{3}&-\frac{1}{3}\end{pmatrix}\begin{pmatrix}3\\1\end{pmatrix}=\frac{10}{3}x_1-\frac{28}{3}x_2+2x_3$$

因此确定稳定点的方程组为

$$\begin{aligned}\nabla_c f&=0\\ g_1(\boldsymbol{X})&=0\\ g_2(\boldsymbol{X})&=0\end{aligned}$$

即

$$\begin{pmatrix} 10 & -28 & 6 \\ 1 & 1 & 3 \\ 5 & 2 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \\ 5 \end{pmatrix}$$

所得到的解为

$$\boldsymbol{X}_0 \approx (0.81,\ 0.35,\ 0.28)$$

用充分条件检查这一稳定点. 由于 x_3 是独立变量, 故从 $\nabla_c f$ 得到

$$\frac{\partial_c^2 f}{\partial_c x_3^2} = \frac{10}{3}\left(\frac{\mathrm{d}x_1}{\mathrm{d}x_3}\right) - \frac{28}{3}\left(\frac{\mathrm{d}x_2}{\mathrm{d}x_3}\right) + 2 = \left(\frac{10}{3}, -\frac{28}{3}\right)\begin{pmatrix} \frac{\mathrm{d}x_1}{\mathrm{d}x_3} \\ \frac{\mathrm{d}x_2}{\mathrm{d}x_3} \end{pmatrix} + 2$$

根据雅可比方法,

$$\begin{pmatrix} \frac{\mathrm{d}x_1}{\mathrm{d}x_3} \\ \frac{\mathrm{d}x_2}{\mathrm{d}x_3} \end{pmatrix} = -\boldsymbol{J}^{-1}\boldsymbol{C} = \begin{pmatrix} \frac{5}{3} \\ -\frac{14}{3} \end{pmatrix}$$

代换得到 $\frac{\partial_c^2 f}{\partial_c x_3^2} = \frac{460}{9} > 0$, 因此 $\boldsymbol{X}_0$ 是极小值点.

雅可比方法中的灵敏度分析 雅可比方法可用来研究约束右端项小量变化对 f 的最优值的影响. 具体来说, 就是 $g_i(\boldsymbol{X}) = 0$ 变成 $g_i(\boldsymbol{X}) = \partial g_i$ 会对 f 的最优值有什么影响. 这一类研究称为**灵敏度分析** (sensitivity analysis), 类似于线性规划中的灵敏度分析 (见第 3 章和第 4 章). 但是, 非线性规划中的灵敏度分析只在极值点的小邻域内有意义. 灵敏度分析的方法有助于对拉格朗日方法的研究.

前面已经说明了

$$\partial f(\boldsymbol{Y}, \boldsymbol{Z}) = \nabla_{\boldsymbol{Y}} f \partial \boldsymbol{Y} + \nabla_{\boldsymbol{Z}} f \partial \boldsymbol{Z}$$
$$\partial \boldsymbol{g} = \boldsymbol{J} \partial \boldsymbol{Y} + \boldsymbol{C} \partial \boldsymbol{Z}$$

由 $\partial \boldsymbol{g} \neq \boldsymbol{0}$, 我们有

$$\partial \boldsymbol{Y} = \boldsymbol{J}^{-1} \partial \boldsymbol{g} - \boldsymbol{J}^{-1} \boldsymbol{C} \partial \boldsymbol{Z}$$

代入 $\lambda \partial f(\boldsymbol{Y}, \boldsymbol{Z})$ 的方程中, 得

$$\partial f(\boldsymbol{Y}, \boldsymbol{Z}) = \nabla_{\boldsymbol{Y}} f \boldsymbol{J}^{-1} \partial \boldsymbol{g} + \nabla_c f \partial \boldsymbol{Z}$$

其中 (根据前面的定义)

$$\nabla_c f = \nabla_{\boldsymbol{Z}} f - \nabla_{\boldsymbol{Y}} f \boldsymbol{J}^{-1} \boldsymbol{C}$$

$\partial f(\boldsymbol{Y}, \boldsymbol{Z})$ 的表达式可用来研究在可行点 $\boldsymbol{X}_0$ 的可行邻域内由小量变化 $\partial \boldsymbol{g}$ 和 $\partial \boldsymbol{Z}$ 所引起的 f 的变化情况.

在极值点 (实际上是任何稳定点)$\boldsymbol{X}_0=(\boldsymbol{Y}_0,\boldsymbol{Z}_0)$ 处约束梯度 $\nabla_c f$ 必为零, 因此有

$$\partial f(\boldsymbol{Y}_0,\boldsymbol{Z}_0)=\nabla_{\boldsymbol{Y}_0} f \boldsymbol{J}^{-1}\partial \boldsymbol{g}(\boldsymbol{Y}_0,\boldsymbol{Z}_0)$$

或

$$\frac{\partial f}{\partial \boldsymbol{g}}=\nabla_{\boldsymbol{Y}_0} f \boldsymbol{J}^{-1}$$

小量变化 $\partial \boldsymbol{g}$ 对 f 的最优值的影响可以通过考察 f 关于 $\boldsymbol{g}$ 的变化率来研究. 这些变化率通常称为**灵敏度系数** (sensitivity coefficient).

例 18.2-3

考虑例 18.2-2 同样的问题. 最优点由 $\boldsymbol{X}_0=(x_{01},x_{02},x_{03})=(0.81,0.35,0.28)$ 给出. 由于 $\boldsymbol{Y}_0=(x_{01},x_{02})$, 故

$$\nabla_{\boldsymbol{Y}_0} f=\left(\frac{\partial f}{\partial x_1},\frac{\partial f}{\partial x_2}\right)=(2x_{01},2x_{02})=(1.62,0.70)$$

进而有

$$\left(\frac{\partial f}{\partial g_1},\frac{\partial f}{\partial g_2}\right)=\nabla_{\boldsymbol{Y}_0} f \boldsymbol{J}^{-1}=(1.62,0.7)\begin{pmatrix}-\frac{2}{3} & \frac{1}{3}\\ \frac{5}{3} & -\frac{1}{3}\end{pmatrix}=(0.086\ 7,0.306\ 7)$$

这意味着对于 $\partial g_1=1$, f 近似增加 0.086 7. 类似地, 对 $\partial g_2=1$, f 近似增加 0.306 7.

雅可比方法用于线性规划问题 考虑下面的线性规划:

$$\begin{aligned}\max\ & z=2x_1+3x_2\\ \text{s.t.}\ & x_1+x_2+x_3 \qquad =5\\ & x_1-x_2 \qquad +x_4=3\\ & x_1,x_2,\ x_3,\ x_4\geqslant 0\end{aligned}$$

为了处理非负性约束 $x_j\geqslant 0$, 使用替换 $x_j=w_j^2$, 这样非负条件就可以不直接表示出来, 原来的问题变成

$$\begin{aligned}\max\ & z=2w_1^2+3w_2^2\\ \text{s.t.}\ & w_1^2+w_2^2+w_3^2 \qquad =5\\ & w_1^2-w_2^2 \qquad +w_4^2=3\end{aligned}$$

为运用雅可比方法, 令

$$\boldsymbol{Y}=(w_1,w_2),\quad \boldsymbol{Z}=(w_3,w_4)$$

(按照线性规划的术语, $\boldsymbol{Y}$ 和 $\boldsymbol{Z}$ 分别对应于基变量和非基变量.) 因此有

$$\boldsymbol{J}=\begin{pmatrix}2w_1 & 2w_2\\ 2w_1 & -2w_2\end{pmatrix},\quad \boldsymbol{J}^{-1}=\begin{pmatrix}\frac{1}{4w_1} & \frac{1}{4w_1}\\ \frac{1}{4w_2} & \frac{-1}{4w_2}\end{pmatrix},\quad w_1,w_2\neq 0$$

$$\boldsymbol{C}=\begin{pmatrix}2w_3 & 0\\ 0 & 2w_4\end{pmatrix},\quad \nabla_{\boldsymbol{Y}}f=(4w_1,6w_2),\quad \nabla_{\boldsymbol{Z}}f=(0,0)$$

由此有

$$\nabla_{\mathrm{c}}f=(0,0)-(4w_1,6w_2)\begin{pmatrix}\frac{1}{4w_1} & \frac{1}{4w_1}\\ \frac{1}{4w_2} & \frac{-1}{4w_2}\end{pmatrix}\begin{pmatrix}2w_3 & 0\\ 0 & 2w_4\end{pmatrix}=(-5w_3,w_4)$$

由 $\nabla_{\mathrm{c}}f=0$ 得到方程的解, 并由问题的约束得到稳定点 ($w_1=2$, $w_2=1$, $w_3=0$, $w_4=0$). Hesse 矩阵为

$$\boldsymbol{H}_{\mathrm{c}}=\begin{pmatrix}\frac{\partial_{\mathrm{c}}^2 f}{\partial_{\mathrm{c}}w_3^2} & \frac{\partial_{\mathrm{c}}^2 f}{\partial_{\mathrm{c}}w_3\partial_{\mathrm{c}}w_4}\\ \frac{\partial_{\mathrm{c}}^2 f}{\partial_{\mathrm{c}}w_3\partial_{\mathrm{c}}w_4} & \frac{\partial_{\mathrm{c}}^2 f}{\partial_{\mathrm{c}}w_4^2}\end{pmatrix}=\begin{pmatrix}-5 & 0\\ 0 & 1\end{pmatrix}$$

由于 $\boldsymbol{H}_{\mathrm{c}}$ 是非定的, 稳定点不是极大值点.

上述解不是最优解的原因在于, $\boldsymbol{Y}$ 的 $\boldsymbol{Z}$ 的具体选择不是最优的. 事实上, 要找到最优解, 我们需要不断改变对 $\boldsymbol{Y}$ 的 $\boldsymbol{Z}$ 的选择, 直到满足充分条件. 这相当于要找到该线性规划解空间的最优极值点. 例如, 考虑 $\boldsymbol{Y}=(w_2,w_4)$, $\boldsymbol{Z}=(w_1,w_3)$, 相应的约束梯度向量变成

$$\nabla_{\mathrm{c}}f=(4w_1,0)-(6w_2,0)\begin{pmatrix}\frac{1}{2w_2} & 0\\ \frac{1}{2w_4} & \frac{1}{2w_4}\end{pmatrix}\begin{pmatrix}2w_1 & 2w_3\\ 2w_1 & 0\end{pmatrix}=(-2w_1,6w_3)$$

相应的稳定点为 $w_1=0,w_2=\sqrt{5},w_3=0,w_4=\sqrt{8}$. 因为

$$\boldsymbol{H}_{\mathrm{c}}=\begin{pmatrix}-2 & 0\\ 0 & -6\end{pmatrix}$$

是负定的, 该解是极大值点.

图 18.5 给出了这一结果的图形解释. 第 1 个解 ($x_1=4,x_2=1$) 不是最优的, 而第 2 个解 ($x_1=0,x_2=5$) 是最优解. 可以验证解空间剩下的两个 (可行) 极点也不是最优的. 事实上, 可以根据充分条件表明, 极点 ($x_1=0,x_2=0$) 是一个极小值点.

灵敏度系数 $\nabla_{\boldsymbol{Y}_0}f\boldsymbol{J}^{-1}$ 当用于线性规划时给出对偶值. 为了在本例中说明这一点, 令 u_1 和 u_2 为相应的对偶变量, 在最优点 ($w_1=0$, $w_2=\sqrt{5}$, $w_3=0$, $w_4=\sqrt{8}$) 处, 这些对偶变量的值为

$$(u_1,u_2)=\nabla_{\boldsymbol{Y}_0}\boldsymbol{J}^{-1}=(6w_2,0)\begin{pmatrix}\frac{1}{2w_2} & 0\\ \frac{1}{2w_4} & \frac{1}{2w_4}\end{pmatrix}=(3,0)$$

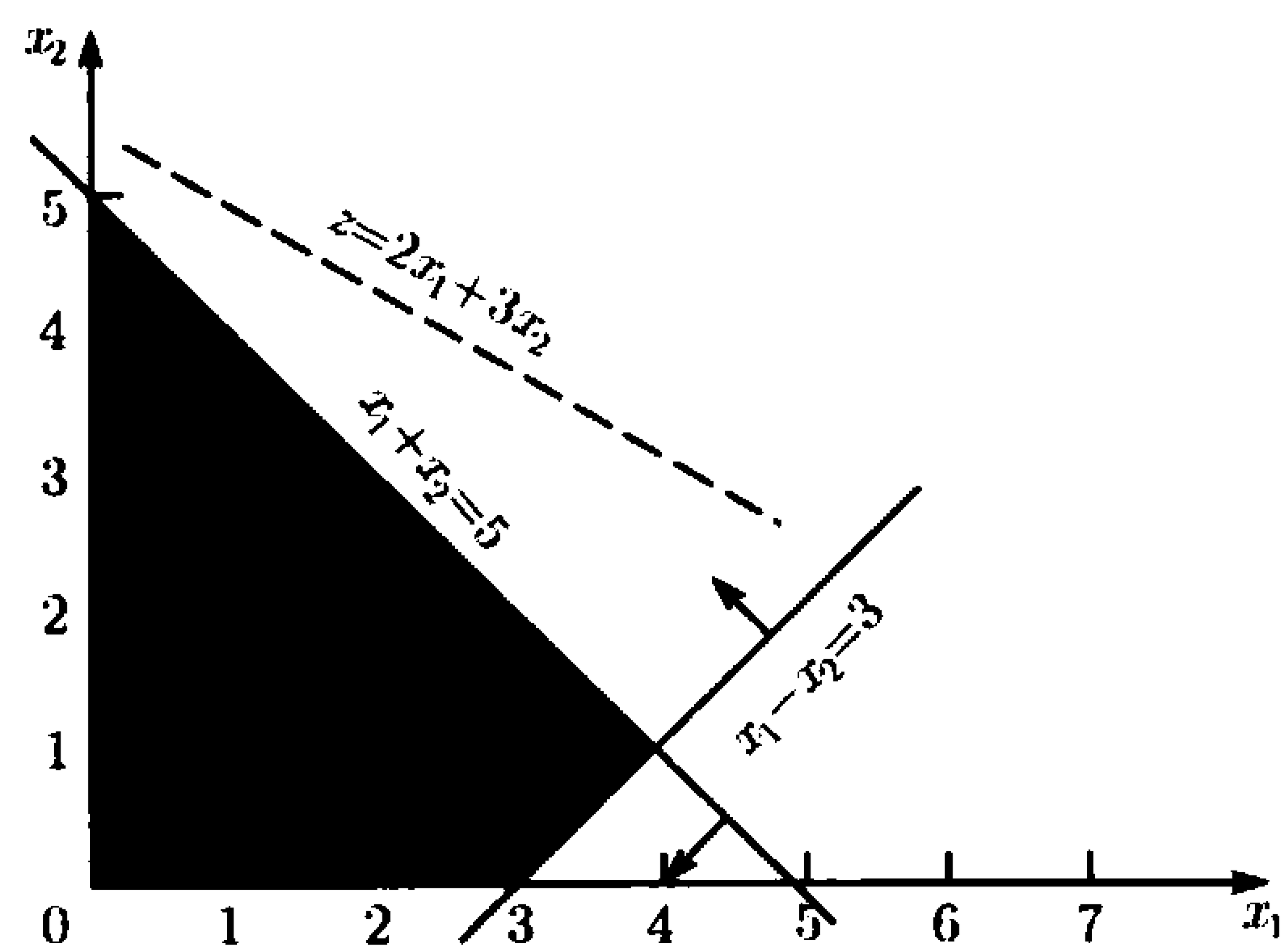


图 18.5　线性规划解空间的极值点

相应的对偶目标值为 $5u_1 + 3u_2 = 15$, 它等于最优原始目标值. 所给出的解还满足对偶约束, 因此是最优可行解. 这表明灵敏度系数和对偶变量是一样的. 事实上它们有着同样的解释.

从雅可比方法用于解线性规划问题中, 可以得出一些一般性的结论. 从数值例子中, 必要条件要求独立变量等于零. 此外, 充分条件表明, Hesse 矩阵是一个对角矩阵. 因此, 对角元素对于极小值点是正的, 对极大值点是负的. 这些现象说明, 必要条件相当于说, 要找到最优解我们只需考虑基本 (可行) 解. 在这种情况下, 独立变量等价于线性规划中的非基变量. 此外, 充分条件表明了 Hesse 矩阵对角元素与单纯形方法中最优性指标 $z_j - c_j$ (见 13.2 节) 之间的密切关系①.

习题 18.2B

1. 假定按下列方法求解例 18.2-2. 首先, 约束中用 x_3 表示 x_1 和 x_2, 然后用所得到的方程表示只有 x_3 的目标函数. 通过求新目标函数关于 x_3 的导数, 可以求出极大值点和极小值点.
   (a) (用 x_3 表示的) 新的目标函数的导数与用雅可比方法得到的导数会有不同吗?
   (b) 所提出的方法与雅可比方法有什么不同?
2. 选定 $\boldsymbol{Y} = (x_2, x_3), \boldsymbol{Z} = (x_1)$, 用雅可比方法求解例 18.2-1.

*3. 用雅可比方法求解:

$$\min\ f(\boldsymbol{X}) = \sum_{i=1}^{n} x_i^2$$

$$\text{s.t.}\ \prod_{i=1}^{n} x_i = C$$

① 关于这些结果对一般线性规划问题的正确性证明, 见 H. Taha and G. Curry, "Classical Derivation of the Necessary and Sufficient Conditions for Optimal Linear Programs," *Operations Research*, Vol. 19, pp. 1045-1049, 1971. 这篇文章证明了, 单纯形方法的关键思想可以从雅可比方法导出.

其中 C 为正常数. 假定约束的右端项变成 $C+\delta$, δ 是一个正小量. 求出 f 最优值相应的变化量.

4. 用雅可比方法求解

$$\begin{aligned} \min \quad & f(\boldsymbol{X}) = 5x_1^2 + x_2^2 + 2x_1x_2 \\ \text{s.t.} \quad & g(\boldsymbol{X}) = x_1x_2 - 10 = 0 \end{aligned}$$

(a) 若约束替换成 $x_1x_2 - 9.99 = 0$, 找出 $f(\boldsymbol{X})$ 最优值的改变量.

(b) 当 $x_1x_2 = 9.99$ 并且 $\partial x_1 = 0.01$, 在可行点 $(2,5)$ 的邻域内找出 $f(\boldsymbol{X})$ 值的改变量.

5. 考虑问题:

$$\begin{aligned} \max \quad & f(\boldsymbol{X}) = x_1^2 + 2x_2^2 + 10x_3^2 + 5x_1x_2 \\ \text{s.t.} \quad & g_1(\boldsymbol{X}) = x_1 + x_2^2 + 3x_2x_3 - 5 = 0 \\ & g_2(\boldsymbol{X}) = x_1^2 + 5x_1x_2 + x_3^2 - 7 = 0 \end{aligned}$$

用雅可比方法在可行点 $(1,1,1)$ 的邻域中找出 $\partial f(\boldsymbol{X})$. 设该邻域由 $\partial g_1 = -0.01$, $\partial g_2 = 0.02$, $\partial x_1 = 0.01$ 确定.

6. 考虑问题

$$\begin{aligned} \min \quad & f(\boldsymbol{X}) = x_1^2 + x_2^2 + x_3^2 + x_4^2 \\ \text{s.t.} \quad & g_1(\boldsymbol{X}) = x_1 + 2x_2 + 3x_3 + 5x_4 - 10 = 0 \\ & g_2(\boldsymbol{X}) = x_1 + 2x_2 + 5x_3 + 6x_4 - 15 = 0 \end{aligned}$$

(a) 说明通过选择 x_3, x_4 作为独立变量, 雅可比方法得不到一个解, 并阐述其原因.

*(b) 用 x_1, x_3 作为独立变量求解该问题, 并用充分条件确定所得到的稳定点的类型.

(c) 确定 (b) 中给出的解的灵敏度系数.

7. 考虑线性规划问题:

$$\begin{aligned} \max \quad & f(\boldsymbol{X}) = \sum_{j=1}^{n} c_j x_j \\ \text{s.t.} \quad & g_i(\boldsymbol{X}) = \sum_{j=1}^{n} a_{ij}x_j - b_j = 0, \quad i = 1,2,\cdots,m \\ & x_j \geqslant 0, \quad j = 1,2,\cdots,n \end{aligned}$$

忽略非负约束, 证明该问题的约束导数 $\nabla_c f(\boldsymbol{X})$ 给出了与由线性规划问题的最优性条件定义的 $\{z_j - c_j\}$ 相同的表达式, 即

$$\{z_j - c_j\} = \{\boldsymbol{C_B B}^{-1}\boldsymbol{P}_j - c_j\}, \quad \text{对于所有的} j$$

约束导数方法能否直接用于求解线性规划问题? 为什么?

拉格朗日方法 在雅可比方法中, 令向量 $\boldsymbol{\lambda}$ 表示灵敏度系数, 即

$$\boldsymbol{\lambda} = \nabla_{\boldsymbol{Y}_0}\boldsymbol{J}^{-1} = \frac{\partial f}{\partial g}$$

因此,

$$\partial f - \boldsymbol{\lambda}\partial g = 0$$

因为 $\frac{\partial f}{\partial g}$ 的计算使得 $\nabla_c f = \boldsymbol{0}$, 所以这个方程满足必要条件. 得到这些方程的一种更加方便的形式, 是求关于所有 x_j 的偏导数. 于是得到

$$\frac{\partial}{\partial x_j}(f - \boldsymbol{\lambda} g) = 0, \quad j = 1, 2, \cdots, n$$

这些方程和约束方程 $g(\boldsymbol{X}) = \boldsymbol{0}$ 一起, 给出满足稳定点必要条件的 $\boldsymbol{X}$ 和 $\boldsymbol{\lambda}$ 值.

以上方法定义了求带有等式约束最优化问题稳定点的拉格朗日方法. 令

$$L(\boldsymbol{X}, \boldsymbol{\lambda}) = f(\boldsymbol{X}) - \boldsymbol{\lambda} g(\boldsymbol{X})$$

函数 L 称为**拉格朗日函数**, 参数$\boldsymbol{\lambda}$称为**拉格朗日乘子**. 按照定义, 这些乘子与雅可比方法中的灵敏度系数具有相同的解释.

方程

$$\frac{\partial L}{\partial \boldsymbol{\lambda}} = 0, \quad \frac{\partial L}{\partial \boldsymbol{X}} = 0$$

给出了 $f(\boldsymbol{X})$ 在约束 $g(\boldsymbol{X}) = \boldsymbol{0}$ 下稳定点的必要条件. 拉格朗日方法的充分条件存在, 但一般计算上非常复杂.

例 18.2-4

考虑例 18.2-2 的问题. 拉格朗日函数是

$$L(\boldsymbol{X}, \boldsymbol{\lambda}) = x_1^2 + x_2^2 + x_3^2 - \lambda_1(x_1 + x_2 + 3x_3 - 2) - \lambda_2(5x_1 + 2x_2 + x_3 - 5)$$

由此得出下列必要条件:

$$\begin{aligned}
\frac{\partial L}{\partial x_1} &= 2x_1 - \lambda_1 - 5\lambda_2 = 0 \\
\frac{\partial L}{\partial x_2} &= 2x_2 - \lambda_1 - 2\lambda_2 = 0 \\
\frac{\partial L}{\partial x_3} &= 2x_3 - 3\lambda_1 - \lambda_2 = 0 \\
\frac{\partial L}{\partial \lambda_1} &= -(x_1 + x_2 + 3x_3 - 2) = 0 \\
\frac{\partial L}{\partial \lambda_2} &= -(5x_1 + 2x_2 + x_3 - 5) = 0
\end{aligned}$$

这些联立方程的解为

$$\boldsymbol{X}_0 = (x_1, x_2, x_3) = (0.804\,3, 0.347\,8, 0.282\,6)$$

$$\boldsymbol{\lambda} = (\lambda_1, \lambda_2) = (0.087\,0, 0.304\,3)$$

这个解同时给出了例 18.2-2 和例 18.2-3 的结果. 拉格朗日乘子 $\boldsymbol{\lambda}$ 的值等于例 18.2-3 得到的灵敏度系数. 这一结果说明, 这些系数与雅可比方法中相关向量 $\boldsymbol{Y}$ 的具体选择无关.

习题 18.2C

1. 分别用雅可比方法和拉格朗日方法求解下面的线性规划问题:

$$\begin{aligned} \max\ & f(\boldsymbol{X}) = 5x_1 + 3x_2 \\ \text{s.t.}\ & g_1(\boldsymbol{X}) = x_1 + 2x_2 + x_3 \qquad - 6 = 0 \\ & g_2(\boldsymbol{X}) = 3x_1 + x_2 \qquad + x_4 - 9 = 0 \\ & x_1, x_2, x_3, x_4 \geqslant 0 \end{aligned}$$

*2. 求以下问题的最优解:

$$\begin{aligned} \min\ & f(\boldsymbol{X}) = x_1^2 + 2x_2 + 10x_3^2 \\ \text{s.t.}\ & g_1(\boldsymbol{X}) = x_1 + x_2^2 + x_3 - 5 = 0 \\ & g_2(\boldsymbol{X}) = x_1 + 5x_2 + x_3 - 7 = 0 \end{aligned}$$

假设 $g_1(\boldsymbol{X}) = 0.01$, $g_2(\boldsymbol{X}) = 0.02$, 求出 $f(\boldsymbol{X})$ 最优值的相应改变值.

3. 用拉格朗日方法求解习题 18.2B 中的第 6 题, 并验证拉格朗日乘子的值与得到的灵敏度系数相同.

18.2.2 不等式约束问题: Karush-Kuhn-Tucker (KKT) 条件[①]

本节将拉格朗日方法推广到不等式约束问题. 本节的主要内容是关于稳定点的一般性 Karush-Kuhn-Tucker(KKT)必要条件, 这些条件在后面要介绍的某些规则下也是充分的.

考虑问题

$$\begin{aligned} \max\ & z = f(\boldsymbol{X}) \\ \text{s.t.}\ & \boldsymbol{g}(\boldsymbol{X}) \leqslant \boldsymbol{0} \end{aligned}$$

用非负松弛变量, 上述不等式约束也可以变换成等式约束. 令 $S_i^2(\geqslant 0)$ 为加到第 i 个约束 $g_i(\boldsymbol{X}) \leqslant 0$ 上的松弛量, 并且定义

$$\boldsymbol{S} = (S_1, S_2, \cdots, S_m)^{\mathrm{T}}, \quad \boldsymbol{S}^2 = (S_1^2, S_2^2, \cdots, S_m^2)^{\mathrm{T}}$$

其中 m 是不等式约束的总个数. 因此拉格朗日函数为

$$L(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\lambda}) = f(\boldsymbol{X}) - \boldsymbol{\lambda}[\boldsymbol{g}(\boldsymbol{X}) + \boldsymbol{S}^2]$$

① 历史上, W. Karush 在 1939 年芝加哥大学的硕士论文中首次提出了 KKT 条件. 这些条件于 1951 年分别由 W. Kuhn 和 A. Tucker 独立发现.

根据约束

$$\boldsymbol{g}(\boldsymbol{X}) \leqslant \boldsymbol{0}$$

对于求最大值 (最小值) 问题的最优性必要条件为 λ 非负 (非正). 验证这一结果的正确性, 只要注意到向量 $\boldsymbol{\lambda}$ 度量了 f 关于 $\boldsymbol{g}$ 的变化率, 即

$$\boldsymbol{\lambda} = \frac{\partial f}{\partial \boldsymbol{g}}$$

在求最大值情况下, 随着约束 $\boldsymbol{g}(\boldsymbol{X}) \leqslant \boldsymbol{0}$ 的右端项从 $\boldsymbol{0}$ 增加到 $\partial \boldsymbol{g}$, 解空间会变得约束更松, 因此 f 不能减少, 意味着 $\boldsymbol{\lambda} \geqslant \boldsymbol{0}$. 对于求最小值, 情况类似, 随着约束的右端项增加, f 不能增加, 这就意味着 $\boldsymbol{\lambda} \leqslant \boldsymbol{0}$. 如果是等式约束, 即 $\boldsymbol{g}(\boldsymbol{X}) = \boldsymbol{0}$, 则 $\boldsymbol{\lambda}$ 无符号限制 (见习题 18.2D 第 2 题).

对 $\boldsymbol{\lambda}$ 的限制只是 KKT 必要条件的一部分, 现在推导剩下的条件.

求 L 关于 $\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\lambda}$ 的偏导数, 我们得到

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{X}} &= \nabla f(\boldsymbol{X}) - \boldsymbol{\lambda} \nabla \boldsymbol{g}(\boldsymbol{X}) = \boldsymbol{0} \\
\frac{\partial L}{\partial S_i} &= -2\lambda_i S_i = 0, \quad i = 1, 2, \cdots, m \\
\frac{\partial L}{\partial \boldsymbol{\lambda}} &= -\left(\boldsymbol{g}(\boldsymbol{X}) + \boldsymbol{S}^2\right) = \boldsymbol{0}
\end{aligned}$$

其中第 2 组方程揭示出下面的结果:

(1) 若 $\lambda_i \neq 0$, 则 $S_i^2 = 0$, 这意味着相应的资源紧缺, 因此完全耗尽 (等式约束).

(2) 若 $S_i^2 > 0$, 则 $\lambda_i = 0$, 这意味着资源 i 充沛, 因此它对 f 的取值没有影响 (即, $\lambda_i = \frac{\partial f}{\partial g_i} = 0$).

从第 2 组和第 3 组方程我们得到

$$\lambda_i g_i(\boldsymbol{X}) = 0, \ i = 1, 2, \cdots, m$$

这个新条件基本上是重复前面的论点, 因为若 $\lambda_i > 0$ 则 $g_i(\boldsymbol{X}) = 0$ 或 $S_i^2 = 0$; 若 $g_i(\boldsymbol{X}) < 0$, 则 $S_i^2 > 0$, 并且 $\lambda_i = 0$.

求最大值问题的 KKT 必要条件总结如下:

$$\begin{aligned}
\boldsymbol{\lambda} &\geqslant \boldsymbol{0} \\
\nabla f(\boldsymbol{X}) - \boldsymbol{\lambda} \nabla \boldsymbol{g}(\boldsymbol{X}) &= \boldsymbol{0} \\
\lambda_i g_i(\boldsymbol{X}) &= 0, \quad i = 1, 2, \cdots, m \\
\boldsymbol{g}(\boldsymbol{X}) &\leqslant \boldsymbol{0}
\end{aligned}$$

这些条件除 $\boldsymbol{\lambda}$ 必须为非正外 (请证明!) 也适用于求最小值情况. 在求最大值和求最小值这两种问题中, 相应于等式约束的拉格朗日乘子没有符号限制.

KKT 条件的充分性 当目标函数和解空间满足某些条件时, Karush-Kuhn-Tucker 必要条件也是充分的. 表 18.1 总结了这些条件.

表 18.1

最优化要求	所需条件	
	目标函数	解空间
求最大值	凹函数	凸集
求最小值	凸函数	凸集

检验函数是凸函数或凹函数比证明解空间是凸集要简单些. 因此, 根据解空间的凸性可以通过检查约束函数的凸性或凹性来得到, 我们给出了一组在实际中方便使用的条件. 为了给出这些条件, 我们定义一般性的非线性规划问题:

$$\begin{aligned} \max\ & z = f(\boldsymbol{X}) \\ \text{s.t.}\ & g_i(\boldsymbol{X}) \leqslant 0,\ i = 1, 2, \cdots, r \\ & g_i(\boldsymbol{X}) \geqslant 0,\ i = r+1, \cdots, p \\ & g_i(\boldsymbol{X}) = 0,\ i = p+1, \cdots, m \end{aligned}$$

$$L(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\lambda}) = f(\boldsymbol{X}) - \sum_{i=1}^{r} \lambda_i \left[g_i(\boldsymbol{X}) + S_i^2\right] - \sum_{i=r+1}^{p} \lambda_i \left[g_i(\boldsymbol{X}) - S_i^2\right] - \sum_{i=p+1}^{m} \lambda_i g_i(\boldsymbol{X})$$

其中 λ_i 是关于约束 i 的拉格朗日乘子. 建立 KKT 条件的充分性的条件列在表 18.2 中.

表 18.2 只表达了表 18.1 中条件的一个子集, 因为不满足表 18.2 条件的一个解空间也可能是凸的.

表 18.2

最优化要求	所需条件			
	$f(\boldsymbol{X})$	$g_i(\boldsymbol{X})$	λ_i	
求最大值	凹函数	凸	$\geqslant 0$	$(1 \leqslant i \leqslant r)$
		凹	$\leqslant 0$	$(r+1 \leqslant i \leqslant p)$
		线性	无限制	$(p+1 \leqslant i \leqslant m)$
求最小值	凸函数	凸	$\leqslant 0$	$(1 \leqslant i \leqslant r)$
		凹	$\geqslant 0$	$(r+1 \leqslant i \leqslant p)$
		线性	无限制	$(p+1 \leqslant i \leqslant m)$

表 18.2 是正确的, 因为所给出的条件确定了求最大值情况下的一个凹拉格朗日函数 $L(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\lambda})$ 和求最小值情况下的一个凸拉格朗日函数 $L(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\lambda})$. 我们只要注意到若 $g_i(x)$ 是凸的, 则当 $\lambda_i \geqslant 0$ 时 $\lambda_i g_i(x)$ 也是凸的, 当 $\lambda_i \leqslant 0$ 时它是凹的. 对所有其他条件可作出类似的解释. 注意, 线性函数既是凸的也是凹的, 还有, 若函数 f 是凹函数, 则 $(-f)$ 是凸的, 反之亦然.

例 18.2-5

考虑以下求最小值问题:

$$
\begin{aligned}
\min\ f(\boldsymbol{X}) &= x_1^2+x_2^2+x_3^2 \\
\text{s.t. } g_1(\boldsymbol{X}) &= 2x_1+x_2-5\leqslant 0 \\
g_2(\boldsymbol{X}) &= x_1+x_3-2\leqslant 0 \\
g_3(\boldsymbol{X}) &= 1-x_1\leqslant 0 \\
g_4(\boldsymbol{X}) &= 2-x_2\leqslant 0 \\
g_5(\boldsymbol{X}) &= -x_3\leqslant 0
\end{aligned}
$$

这是一个求最小值问题, 所以 $\boldsymbol{\lambda}\leqslant\boldsymbol{0}$, 因此 KKT 条件为

$$
\begin{aligned}
(\lambda_1,\lambda_2,\lambda_3,\lambda_4,\lambda_5)&\leqslant\boldsymbol{0} \\
(2x_1,2x_2,2x_3)-(\lambda_1,\lambda_2,\lambda_3,\lambda_4,\lambda_5)\begin{pmatrix}2&1&0\\1&0&1\\-1&0&0\\0&-1&0\\0&0&-1\end{pmatrix}&=\boldsymbol{0} \\
\lambda_1g_1=\lambda_2g_2=\cdots=\lambda_5g_5&=0 \\
\boldsymbol{g}(\boldsymbol{X})&\leqslant\boldsymbol{0}
\end{aligned}
$$

把这些条件简化成

$$
\begin{aligned}
\lambda_1,\lambda_2,\lambda_3,\lambda_4,\lambda_5&\leqslant 0 \\
2x_1-2\lambda_1-\lambda_2+\lambda_3&=0 \\
2x_2-\lambda_1+\lambda_4&=0 \\
2x_3-\lambda_2+\lambda_5&=0 \\
\lambda_1(2x_1+x_2-5)&=0 \\
\lambda_2(x_1+x_3-2)&=0 \\
\lambda_3(1-x_1)&=0 \\
\lambda_4(2-x_2)&=0 \\
\lambda_5x_3&=0 \\
2x_1+x_2&\leqslant 5 \\
x_1+x_3&\leqslant 2 \\
x_1\geqslant 1,x_2\geqslant 2,x_3&\geqslant 0
\end{aligned}
$$

求得的解为 $x_1 = 1$, $x_2 = 2$, $x_3 = 0$, $\lambda_1 = \lambda_2 = \lambda_5 = 0$, $\lambda_3 = -2$, $\lambda_4 = -4$. 因为 $f(\boldsymbol{X})$ 和解空间 $\boldsymbol{g}(\boldsymbol{X}) \leqslant \boldsymbol{0}$ 都是凸的, 所以 $L(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\lambda})$ 必是凸的, 且得到的稳定点为一个全局约束最小值点. KKT 条件成为了第 19 章设计非线性规划算法的核心基础.

习题 18.2D

1. 考虑问题

$$\begin{aligned} \max\ & f(\boldsymbol{X}) \\ \text{s.t.}\ & \boldsymbol{g}(\boldsymbol{X}) \geqslant \boldsymbol{0} \end{aligned}$$

说明除了拉格朗日乘子 $\boldsymbol{\lambda}$ 非正外, KKT 条件与 18.2.2 节中的相同.

2. 考虑以下问题:

$$\begin{aligned} \max\ & f(\boldsymbol{X}) \\ \text{s.t.}\ & \boldsymbol{g}(\boldsymbol{X}) = \boldsymbol{0} \end{aligned}$$

证明 KKT 条件为

$$\begin{aligned} \nabla f(\boldsymbol{X}) - \boldsymbol{\lambda} \nabla \boldsymbol{g}(\boldsymbol{X}) &= \boldsymbol{0} \\ \boldsymbol{g}(\boldsymbol{X}) &= \boldsymbol{0} \\ \boldsymbol{\lambda}\ \text{无符号限制} \end{aligned}$$

3. 写出下列问题的 KKT 必要条件.

(a)
$$\begin{aligned} \max\ & f(\boldsymbol{X}) = x_1^3 - x_2^2 + x_1 x_3^2 \\ \text{s.t.}\ & x_1 + x_2^2 + x_3 = 5 \\ & 5x_1^2 - x_2^2 - x_3 \geqslant 2 \\ & x_1,\ x_2,\ x_3 \geqslant 0 \end{aligned}$$

(b)
$$\begin{aligned} \min\ & f(\boldsymbol{X}) = x_1^4 + x_2^2 + 5x_1 x_2 x_3 \\ \text{s.t.}\ & x_1^2 - x_2^2 + x_3^3 \leqslant 10 \\ & x_1^3 + x_2^2 + 4x_3^2 \geqslant 20 \end{aligned}$$

4. 考虑问题

$$\begin{aligned} \max\ & f(\boldsymbol{X}) \\ \text{s.t.}\ & \boldsymbol{g}(\boldsymbol{X}) = \boldsymbol{0} \end{aligned}$$

设 $f(\boldsymbol{X})$ 是凹函数, 且 $g_i(\boldsymbol{X})(i = 1, 2, \cdots, m)$ 为线性函数, 证明 KKT 必要条件也是充分的. 当 $g_i(\boldsymbol{X})$ 对所有的 i 为非线性凸函数时, 这个结果还正确吗? 为什么?

5. 考虑问题

$$\begin{aligned} \max\ & f(\boldsymbol{X}) \\ \text{s.t.}\ & g_1(\boldsymbol{X}) \geqslant 0,\ g_2(\boldsymbol{X}) = 0,\ g_3(\boldsymbol{X}) \leqslant 0 \end{aligned}$$

推出 KKT 条件, 并给出保证这些条件也是充分的前提约定.

参 考 文 献

Bazarra, M., H. Sherali, and C. Shetty, *Nonlinear Programming Theory and Algorithms,* 2nd ed., Wiley, New York, 1993.

Beightler, C., D. Phillips, and D. Wilde, *Foundations of Optimization*, 2nd ed., Prentice Hall, NJ, 1979.

Rardin, R., *Optimization in Operations Research*, Prentice Hall, NJ, 1998.

第19章　非线性规划算法

本章导读　非线性规划的求解方法一般分成*直接方法*和*间接方法*. 比如, 梯度算法就是一种直接方法, 它通过跟踪目标函数的最速上升 (下降) 率来找到问题的最大值 (最小值). 间接方法是用某个辅助问题代替原来的问题, 并从辅助问题求出最优值. 间接方法用于求解二次规划、可分离规划和随机规划等.

本章包括 9 个例题、1 个 AMPL 模型、2 个 Excel 规划求解模型和 24 个节后习题. AMPL/Excel/Solver/TORA 程序放在文件夹 ch19Files 下.

19.1　无约束算法

本节介绍两种求无约束问题的算法: *直接搜索*算法和*梯度*算法.

19.1.1　直接搜索方法

直接搜索方法主要用于严格单峰的一元函数. 虽然这种情况看上去很简单, 但 19.1.2 节表明, 一元函数的最优化是更一般的多元算法设计的关键.

直接搜索方法的思路是, 找出已知包含最优解点的**不确定区间** (interval of uncertainty). 该方法通过迭代逐渐缩小不确定区间的宽度, 以任何所需的精度来近似确定最优点的位置.

本节介绍两种密切相关的算法: **二分搜索法** (dichotomous) 和**黄金分割搜索法** (golden section). 这两种算法都要在已知包含最优值点 x^* 的区间 $a \leqslant x \leqslant b$ 内找出单峰函数 $f(x)$ 上的最大值. 这两种方法都从 $I_0 = (a, b)$ 开始, 它代表初始不确定区间.

一般步骤 i　令 $I_{i-1} = (x_L, x_R)$ 为当前的不确定区间 (在迭代 0 步, $x_L = a, x_R = b$). 接下来, 按以下方法找出 x_1 和 x_2:

二　分　法	黄金分割法
$x_1 = \frac{1}{2}(x_R + x_L - \Delta)$	$x_1 = x_R - \left(\frac{\sqrt{5}-1}{2}\right)(x_R - x_L)$
$x_2 = \frac{1}{2}(x_R + x_L + \Delta)$	$x_2 = x_L + \left(\frac{\sqrt{5}-1}{2}\right)(x_R - x_L)$

选择 x_1 和 x_2, 保证

$$x_L < x_1 < x_2 < x_R$$

下一个不确定区间 I_i 按下列步骤确定.

(1) 若 $f(x_1) > f(x_2)$, 则 $x_L < x^* < x_2$. 令 $x_R = x_2$, 并设置 $I_i = (x_L, x_2)$[见图 19.1(a)].

(2) 若 $f(x_1) < f(x_2)$, 则 $x_1 < x^* < x_R$. 令 $x_L = x_1$, 并设置 $I_i = (x_1, x_R)$[见图 19.1(b)].

(3) 若 $f(x_1) = f(x_2)$, 则 $x_1 < x^* < x_2$. 令 $x_L = x_1$, $x_R = x_2$, 并设置 $I_i = (x_1, x_2)$.

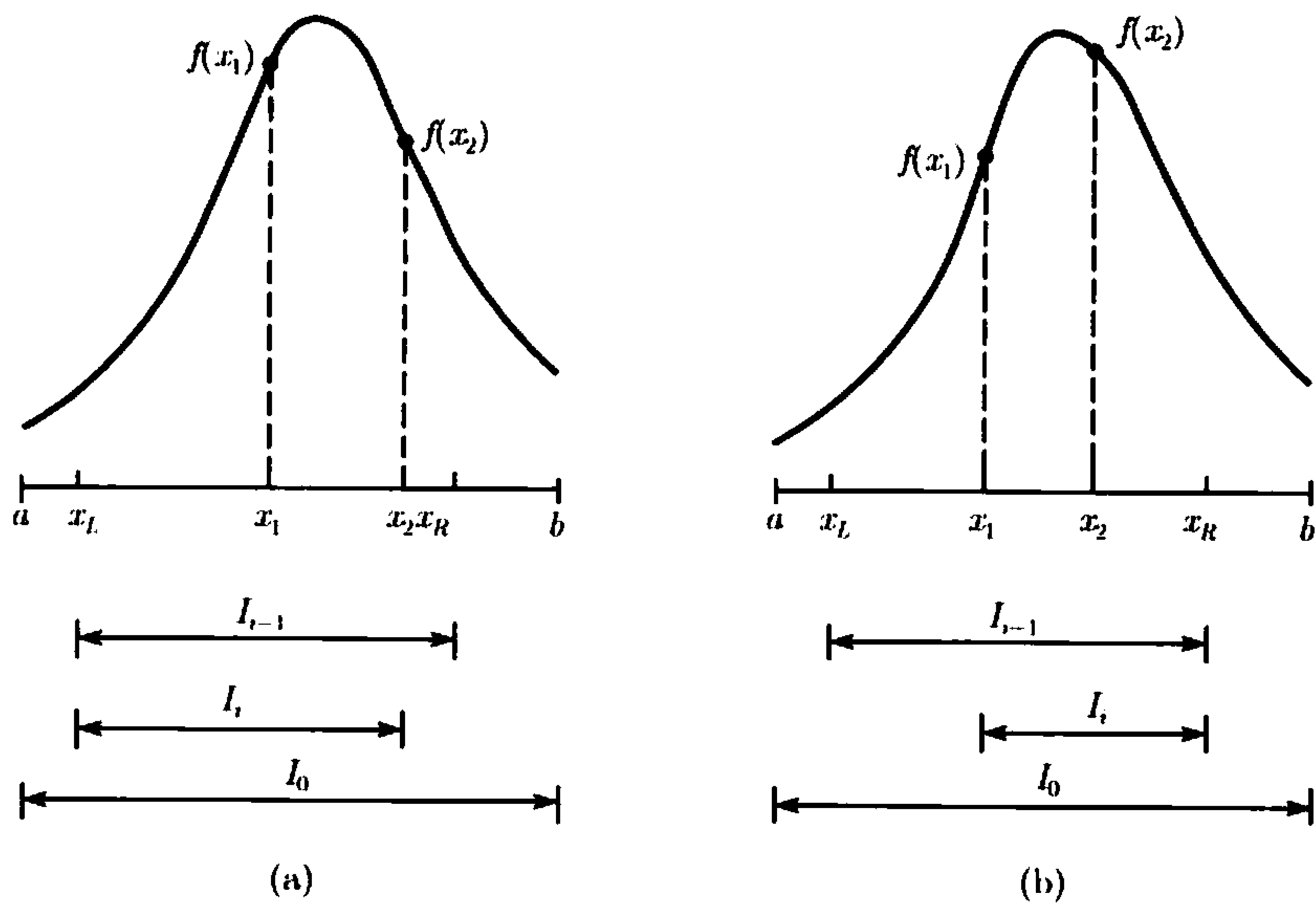


图 19.1 二分搜索法/黄金分割搜索法的一般步骤的说明

确定 x_1 和 x_2 的方法保证了 $I_{i+1} < I_i$, 如下面将要说明的. 若 $I_k \leqslant \Delta$, 算法在第 k 次迭代终止, 其中 Δ 是一个用户指定的精度阈值.

在二分搜索方法中, x_1 和 x_2 的值关于当前不确定区间的中点对称, 这意味着

$$I_{i+1} = 0.5(I_i + \Delta)$$

重复运用这一算法就保证了, 不确定区间的长度趋向于所需的精度 Δ.

黄金分割法更是采用了这一思想. 我们注意到, 二分法的每次迭代要计算两个值 $f(x_1)$ 和 $f(x_2)$, 但最后又放弃了一个. 黄金分割搜索法希望通过在接着的下一次迭代中将放弃的这个值利用起来, 以节省计算量.

对 $0 < \alpha < 1$ 定义

$$x_1 = x_R - \alpha(x_R - x_L)$$
$$x_2 = x_L + \alpha(x_R - x_L)$$

则第 i 步迭代的不确定区间 I_i 等于 (x_L, x_2) 或 (x_1, x_R). 考虑 $I_i = (x_L, x_2)$ 的情况, 意味着 x_1 包含在 I_i 中. 在第 $i+1$ 次迭代中, 选择 x_2 等于第 i 次迭代中的 x_1, 我们得到下面的公式:

$$x_2 \text{ (迭代 } i+1) = x_1 \text{ (迭代 } i)$$

代换得到

$$x_L + \alpha[x_2 \text{ (迭代 } i) - x_L] = x_R - \alpha(x_R - x_L)$$

或

$$x_L + \alpha[x_L + \alpha(x_R - x_L) - x_L] = x_R - \alpha(x_R - x_L)$$

最后简化为

$$\alpha^2 + \alpha - 1 = 0$$

从这个式子得出 $\alpha = \frac{-1\pm\sqrt{5}}{2}$. 由 $0 < \alpha < 1$, 我们选择正根 $\alpha = \frac{-1+\sqrt{5}}{2} \approx 0.618$.

黄金分割法的设计保证每次迭代的不确定区间减小为上一次迭代的 α 倍, 即

$$I_{i+1} = \alpha I_i$$

黄金分割法比二分法收敛更快, 因为在二分法中随着 $I \to \Delta$, 不确定区间的缩小逐渐变慢. 此外, 黄金分割法的每次迭代只需要一半的计算量, 因为该算法总是重复利用上一次迭代的结果.

例 19.1-1

$$\max \quad f(x) = \begin{cases} 3x, & 0 \leqslant x \leqslant 2 \\ \dfrac{1}{3}(-x+20), & 2 \leqslant x \leqslant 3 \end{cases}$$

$f(x)$ 的最大值在 $x = 2$ 取得. 下表展示了分别用二分法和黄金分割法进行第 1 步迭代和第 2 步迭代的计算过程和结果, 假设 $\Delta = 0.1$.

二 分 法	黄金分割法
第 1 步迭代	第 1 步迭代
$I_0 = (0,3) \equiv (x_L, x_R)$	$I_0 = (0,3) \equiv (x_L, x_R)$
$x_1 = 0 + 0.5(3 - 0 - 0.1) = 1.45$	$x_1 = 3 - 0.618(3 - 0) = 1.146$
$f(x_1) = 4.35$	$f(x_1) = 3.438$
$x_2 = 0 + 0.5(3 - 0 + 0.1) = 1.55$	$x_2 = 0 + 0.618(3 - 0) = 1.854$
$f(x_2) = 4.65$	$f(x_2) = 5.562$
$f(x_2) > f(x_1) \Rightarrow x_L = 1.45$	$f(x_2) > f(x_1) \Rightarrow x_L = 1.146$
$I_1 = (1.45, 3)$	$I_1 = (1.146, 3)$
第 2 步迭代	第 2 步迭代
$I_1 = (1.45, 3) \equiv (x_L, x_R)$	$I_1 = (1.146, 3) \equiv (x_L, x_R)$
$x_1 = 1.45 + 0.5(3 - 1.45 - 0.1) = 2.175$	$x_1 =$ 第 1 步迭代的 $x_2 = 1.854$,
$f(x_1) = 5.942$	$f(x_1) = 5.562$
$x_2 = \frac{3+1.45+0.1}{2} = 2.275$	$x_2 = 1.146 + 0.618(3 - 1.146) = 2.292$
$f(x_2) = 5.908$	$f(x_2) = 5.903$
$f(x_1) > f(x_2) \Rightarrow x_R = 2.275$	$f(x_2) > f(x_1) \Rightarrow x_L = 1.854$
$I_2 = (1.45, 2.275)$	$I_1 = (1.854, 3)$

继续同样的迭代, 不确定区间最终缩小到所需的 Δ 允许范围.

Excel 程序

这两种方法都可以用程序 excelIDiGold.xls 计算. 输入数据包括 $f(x)$, a, b 和 Δ. 函数 $f(x)$ 在单元 E3 中输入为

```
= IF(C3<=2,3*C3,(-C3+20)/3)
```

单元格 C3 起着 $f(x)$ 中 x 的作用. 在单元 B4 和 D4 中输入区间边界 a 和 b, 表示 $f(x)$ 的可允许的搜索范围, 另外, 区间长度允许上限 Δ 输入到单元 B3. 通过在单元 D5(二分法) 或者在 F5(黄金分割法) 中输入字符 x 来选择搜索方法.

图 19.2 对这两种方法进行了比较. 黄金分割法只需要二分法一半的迭代次数, 此外如前面所解释的, 每次迭代只需要一半的计算量.

Dichotomous/Golden Section Search					
Input data: Type f(C3) in E3, where C3 represents x in f(x)					
Δ =	0.1	C3		#VALUE!	
Minimum x =	0	Maximum x =	3		
Solution:	Enter x to select>	Dichotomous	x	GoldenSection	
x* =	2.04001	f(x*) =	5.97002		
Calculations:					
xL	xR	x1	x2	f(x1)	f(x2)
0.000000	3.000000	1.450000	1.550000	4.350000	4.650000
1.450000	3.000000	2.175000	2.275000	5.941667	5.908333
1.450000	2.275000	1.812500	1.912500	5.437500	5.737500
1.812500	2.275000	1.993750	2.093750	5.981250	5.968750
1.812500	2.093750	1.903125	2.003125	5.709375	5.998958
1.903125	2.093750	1.948438	2.048438	5.845313	5.983854
1.948438	2.093750	1.971094	2.071094	5.913281	5.976302
1.971094	2.093750	1.982422	2.082422	5.947266	5.972526
1.982422	2.093750	1.988086	2.088086	5.964258	5.970638
1.988086	2.093750	1.990918	2.090918	5.972754	5.969694
1.988086	2.090918	1.989502	2.089502	5.968506	5.970166
1.989502	2.090918	1.990210	2.090210	5.970630	5.969930
1.989502	2.090210	1.989856	2.089856	5.969568	5.970048
1.989856	2.090210	1.990033	2.090033	5.970099	5.969989
1.989856	2.090033	1.989944	2.089944	5.969833	5.970019
1.989944	2.090033	1.989989	2.089989	5.969966	5.970004
1.989989	2.090033	1.990011	2.090011	5.970033	5.969996
1.989989	2.090011	1.990000	2.090000	5.969999	5.970000
1.990000	2.090011	1.990006	2.090005	5.970016	5.969998
1.990000	2.090005	1.990003	2.090003	5.970008	5.969999

Dichotomous/Golden Section Search					
Input data: Type f(C3) in E3, where C3 represents x in f(x)					
Δ =	0.1	C3		#VALUE!	
Minimum x =	0	Maximum x =	3		
Solution:	Enter x to select>	Dichotomous		GoldenSection	x
x* =	2.00909	f(x*) =	5.99290		
Calculations:					
xL	xR	x1	x2	f(x1)	f(x2)
0.000000	3.000000	1.145898	1.854102	3.437694	5.562306
1.145898	3.000000	1.854102	2.291796	5.562306	5.902735
1.854102	3.000000	2.291796	2.562306	5.902735	5.812565
1.854102	2.562306	2.124612	2.291796	5.958463	5.902735
1.854102	2.291796	2.021286	2.124612	5.992905	5.958463
1.854102	2.124612	1.957428	2.021286	5.872283	5.992905
1.957428	2.124612	2.021286	2.060753	5.992905	5.979749
1.957428	2.060753	1.996894	2.021286	5.990683	5.992905
1.996894	2.060753	2.021286	2.036361	5.992905	5.987880

图 19.2 用二分搜索法和黄金分割搜索法求解例 19.1-1 的 Excel 输出 (文件 excelDiGold.xls)

习题 19.1A

1. 令 $\Delta = 0.01$, 用 Excel 程序 excelDiGold.xls 求解例 19.1-1. 将计算量和精度与图 19.2 的结果进行比较.
2. 用二分法求下列函数的极大值点, 设 $\Delta = 0.05$.

(a) $f(x) = \frac{1}{|(x-3)^3|}$, $2 \leqslant x \leqslant 4$　　(b) $f(x) = x\cos x$, $0 \leqslant x \leqslant \pi$

*(c) $f(x) = x\sin \pi x$, $1.5 \leqslant x \leqslant 2.5$　　(d) $f(x) = -(x-3)^2$, $2 \leqslant x \leqslant 4$

*(e) $f(x) = \begin{cases} 4x, & 0 \leqslant x \leqslant 2 \\ 4-x, & 2 \leqslant x \leqslant 4 \end{cases}$

19.1.2 梯度方法

本节介绍一种二次连续可微函数的最优化方法, 其思路是在函数的梯度方向产生一系列点.

18.1.2 节中介绍的 Newton-Raphson 方法就是一种用来求解方程组的梯度方法. 本节介绍另一种技术, 称为**最速下降** (steepest ascent) 法.

梯度方法终止于梯度向量为零的点. 这只是最优性的必要条件, 而且除非事先知道 $f(\boldsymbol{x})$ 是凸的或凹的, 否则最优性不能保证.

假设求 $f(\boldsymbol{X})$ 的最大值. 令 $\boldsymbol{X}_0$ 为开始计算的初始点, 定义 $\nabla f(\boldsymbol{X}_k)$ 为 f 在 $\boldsymbol{X}_k$ 点的梯度. 方法的原理是求出某个特定的路径 p, 沿着这条路经, 在某个给定的点处 $\frac{\partial f}{\partial p}$ 达到最大. 要得到这一结果, 只要选择连续的点 $\boldsymbol{X}_k$ 和 $\boldsymbol{X}_{k+1}$, 使得

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k + r_k \nabla f(\boldsymbol{X}_k)$$

其中 r_k 是在 $\boldsymbol{X}_k$ 点的**最优步长** (step size).

确定步长 r_k 使得下一个点 $\boldsymbol{X}_{k+1}$ 能让 f 得到最大的改进, 这等价于求 $r = r_k$ 使得函数

$$h(r) = f[\boldsymbol{X}_k + r\nabla f(\boldsymbol{X}_k)]$$

取得最大值. 因为 $h(r)$ 是一个一元函数, 故当 $h(r)$ 是严格单峰时, 可以用 19.1.1 节中的搜索方法来找到最优解.

上述方法当两个连续的试验点 $\boldsymbol{X}_k$ 和 $\boldsymbol{X}_{k+1}$ 近似相等时停止, 这等价于

$$r_k \nabla f(\boldsymbol{X}_k) \approx \boldsymbol{0}$$

由 $r_k \neq 0$, 在 $\boldsymbol{X}_k$ 点满足必要条件 $\nabla f(\boldsymbol{X}_k) = \boldsymbol{0}$.

例 19.1-2

考虑下面的问题:

$$\max \quad f(x_1, x_2) = 4x_1 + 6x_2 - 2x_1^2 - 2x_1x_2 - 2x_2^2$$

精确最优解点为 $(x_1^*, x_2^*) = (\frac{1}{3}, \frac{4}{3})$.

我们说明如何用最速下降法求解这个问题. 给出 f 的梯度如下:

$$\nabla f(\boldsymbol{X}) = (4 - 4x_1 - 2x_2, 6 - 2x_1 - 4x_2)$$

该函数的二次性质决定了任何两个连续点上的梯度是正交的 (互相垂直).

假设我们从初始点 $\boldsymbol{X}_0 = (1,1)$ 开始迭代. 图 19.3 显示了连续的解点.

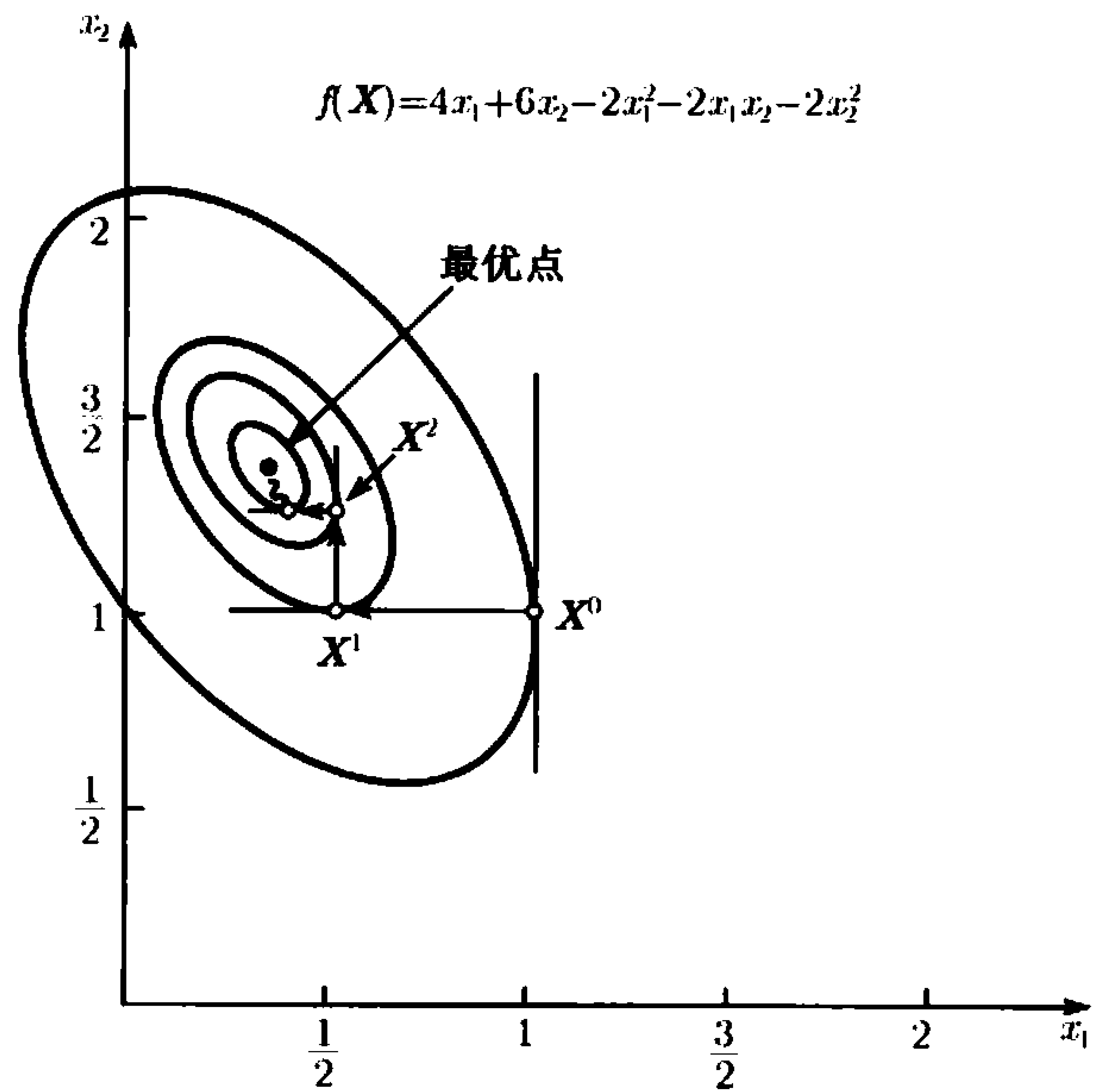


图 19.3　用最速下降法求解例 19.1-2

迭代 1

$$\nabla f(\boldsymbol{X}_0) = (-2, 0)$$

要得到下一个点 $\boldsymbol{X}_1$, 考虑到

$$\boldsymbol{X} = (1,1) + r(-2,0) = (1-2r, 1)$$

因此,

$$h(r) = f(1-2r, 1) = -2(1-2r)^2 + 2(1-2r) + 4$$

用第 18 章中经典的必要条件求出最优步长 (也可以用 19.1.1 节中的搜索算法求出最优点). $h(r)$ 的最大值点是 $r_1 = \frac{1}{4}$, 这得到下一个解点 $\boldsymbol{X}_1 = (\frac{1}{2}, 1)$.

迭代 2

$$\nabla f(\boldsymbol{X}_1) = (0,1)$$

$$\boldsymbol{X} = \left(\frac{1}{2}, 1\right) + r(0,1) = \left(\frac{1}{2}, 1+r\right)$$

$$h(r) = -2(1+r)^2 + 5(1+r) + \frac{3}{2}$$

因此, $r_2=\frac{1}{4}$, $\boldsymbol{X}_2=(\frac{1}{2},\frac{5}{4})$.

迭代 3

$$\nabla f(\boldsymbol{X}_2)=(-\frac{1}{2},0)$$

$$\boldsymbol{X}=\left(\frac{1}{2},\frac{5}{4}\right)+r\left(-\frac{1}{2},0\right)=\left(\frac{1-r}{2},\frac{5}{4}\right)$$

$$h(r)=-\frac{1}{2}(1-r)^2+\frac{3}{4}(1-r)+\frac{35}{8}$$

因此, $r_3=\frac{1}{4}$, $\boldsymbol{X}_3=(\frac{3}{8},\frac{5}{4})$.

迭代 4

$$\nabla f(\boldsymbol{X}_3)=\left(0,\frac{1}{4}\right)$$

$$\boldsymbol{X}=\left(\frac{3}{8},\frac{5}{4}\right)+r\left(0,\frac{1}{4}\right)=\left(\frac{3}{8},\frac{5+r}{4}\right)$$

$$h(r)=-\frac{1}{8}(5+r)^2+\frac{21}{16}(5+r)+\frac{39}{32}$$

得到, $r_4=\frac{1}{4}$, $\boldsymbol{X}_4=(\frac{3}{8},\frac{21}{16})$.

迭代 5

$$\nabla f(\boldsymbol{X}_4)=\left(-\frac{1}{8},0\right)$$

$$\boldsymbol{X}=\left(\frac{3}{8},\frac{21}{16}\right)+r\left(-\frac{1}{8},0\right)=\left(\frac{3-r}{8},\frac{21}{16}\right)$$

$$h(r)=-\frac{1}{32}(3-r)^2+\frac{11}{64}(3-r)+\frac{567}{128}$$

因此, $r_5=\frac{1}{4}$, $\boldsymbol{X}_5=(\frac{11}{32},\frac{21}{16})$.

迭代 6

$$\nabla f(\boldsymbol{X}_5)=\left(0,\frac{1}{16}\right)$$

由于 $\nabla f(\boldsymbol{X}_5)\approx\boldsymbol{0}$, 迭代过程可在此点终止, 近似最大值点由 $\boldsymbol{X}_5=(0.343\,8,1.312\,5)$ 给出. 精确最优点为 $\boldsymbol{X}^*=(0.333\,3,1.333\,3)$.

习题 19.1B

*1. 请说明当用于严格凹二次函数时, Newton-Raphson 方法 (18.1.2 节) 通常会在一步内收敛. 用该方法求下面函数的最大值:

$$f(\boldsymbol{X})=4x_1+6x_2-2x_1^2-2x_1x_2-2x_2^2$$

2. 对于下列每个问题, 用最速下降 (上升) 法最多迭代 5 步. 假设在每种情况下 $\boldsymbol{X}^0=\boldsymbol{0}$.

(a) min $f(\boldsymbol{X}) = (x_2 - x_1^2)^2 + (1 - x_1)$

(b) max $f(\boldsymbol{X}) = \boldsymbol{c}\boldsymbol{X} + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}$

其中

$$\boldsymbol{c} = (1,3,5), \qquad \boldsymbol{A} = \begin{pmatrix} -5 & -3 & -\frac{1}{2} \\ -3 & -2 & 0 \\ -\frac{1}{2} & 0 & -\frac{1}{2} \end{pmatrix}$$

(c) min $f(\boldsymbol{X}) = x_1 - x_2 + x_1^2 - x_1x_2$

19.2 约束算法

一般的约束非线性规划问题定义为

$$\begin{aligned} \max \text{ 或 } \min \quad & z = f(\boldsymbol{X}) \\ \text{s.t.} \quad & g(\boldsymbol{X}) \leqslant \boldsymbol{0} \end{aligned}$$

非负性条件 $\boldsymbol{X} \geqslant \boldsymbol{0}$ 是约束的一部分. 还有, 函数 $f(\boldsymbol{X})$ 和 $g(\boldsymbol{X})$ 至少有一个是非线性的, 并且所有函数都是连续可微的.

由于非线性函数的不确定行为, 所以不存在单一的能处理所有一般性的非线性模型算法. 可用于这类问题的最一般性结果可能就是 KKT 条件 (见 18.2.2 节), 表 18.2 表明, 除非 $f(\boldsymbol{X})$ 和 $g(\boldsymbol{X})$ 具有良好的性质 (根据凸性和凹性条件), 否则 KKT 条件一般只是最优性的必要条件.

本节介绍几种算法, 一般可分为**间接算法**和**直接算法**. 间接算法通过处理从原问题中得到的若干个线性规划来求解非线性问题, 而直接算法则处理原问题.

本节介绍的间接算法包括可分离规划、二次规划和机会约束规划. 直接算法包括线性组合方法, 并简要讨论 SUMT 方法, 即序贯无约束最大化方法. 本章参考文献中还列出了其他一些重要的非线性规划技术.

19.2.1 可分离规划

函数 $f(x_1, x_2, \cdots, x_n)$ 称为**可分离的** (separable), 如果这个函数可表示为 n 个一元函数之和, 即

$$f(x_1, x_2, \cdots, x_n) = f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n)$$

例如, 线性函数

$$h(x_1, x_2, \cdots, x_n) = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

是可分离的 (系数 $a_i, i = 1, 2, \cdots, n$, 为常数). 反过来, 函数

$$h(x_1, x_2, x_3) = x_1^2 + x_1 \sin(x_2 + x_3) + x_2 \mathrm{e}^{x_3}$$

不是可分离的.

某些函数不是直接可分离的，但通过适当的代换就能变成可分离函数. 例如，考虑求 $z = x_1x_2$ 的最大值，令 $y = x_1x_2$，则 $\ln y = \ln x_1 + \ln x_2$，这个问题就变成了

$$\begin{aligned} \max \quad & z = y \\ \text{s.t.} \quad & \ln y = \ln x_1 + \ln x_2 \end{aligned}$$

它是可分离的. 上述代换假定了 x_1 和 x_2 为正的变量，因为对数函数对非正值没有定义.

x_1 和 x_2 可取零值的情况 (即 $x_1, x_2 \geqslant 0$) 可按以下方式处理. 令 δ_1, δ_2 为正常数，定义

$$\begin{aligned} w_1 &= x_1 + \delta_1 \\ w_2 &= x_2 + \delta_2 \end{aligned}$$

在代换

$$x_1x_2 = w_1w_2 - \delta_2w_1 - \delta_1w_2 + \delta_1\delta_2$$

中，新变量 w_1 和 w_2 是严格正的. 令 $y = w_1w_2$，该问题可表示成

$$\begin{aligned} \max \quad & z = y - \delta_2w_1 - \delta_1w_2 + \delta_1\delta_2 \\ \text{s.t.} \quad & \ln y = \ln w_1 + \ln w_2 \\ & w_1 \geqslant \delta_1, \ w_2 \geqslant \delta_2 \end{aligned}$$

它是可分离的.

本节说明，任何可分离问题都可以用线性近似和线性规划的单纯形方法得到一个近似解. 一元函数 $f(x)$ 可以用混合整数规划 (第 8 章) 被分段线性函数近似. 假设要在区间 $[a,b]$ 上近似 $f(x)$，定义 $a_k, k = 1,2,\cdots,K$，为 x 轴上的第 k 个间断点，使得 $a_1 < a_2 < \cdots < a_k$，a_1 和 a_k 恰好是指定区间的端点 a 和 b. 这样，$f(x)$ 被近似如下：

$$\begin{aligned} f(x) &\approx \sum_{k=1}^{K} f(a_k)w_k \\ x &= \sum_{k=1}^{K} a_kw_k \end{aligned}$$

其中 w_k 是关于第 k 个间断点的非负权值，使得

$$\sum_{k=1}^{K} w_k = 1, \quad w_k \geqslant 0, \quad k = 1,2,\cdots,K$$

由下面两个条件, 混合整数规划保证这个近似是合理的.

(1) 最多有两个 w_k 是正的.

(2) 若 w_k 为正, 则仅有一个邻近的 w_{k+1} 或者 w_{k-1} 可取正值.

为了说明如何满足上述条件, 考虑可分离问题

$$\max\ (\text{或}\ \min)\quad z=\sum_{j=1}^{n} f_j(x_j)$$

$$\text{s.t.}\quad \sum_{j=1}^{n} g_{ij}(x_j)\leqslant b_i,\quad i=1,2,\cdots,m$$

这一问题可用一混合整数规划近似如下. 令[①]

$$\left.\begin{array}{l} a_{jk}=\text{变量 } x_j \text{ 的间断点 } k \\ w_{jk}=\text{变量 } x_j \text{ 的间断点 } k \text{ 的权值} \end{array}\right\} k=1,2,\cdots,K_j,\ j=1,2,\cdots,n$$

则等价的混合问题为

$$\max\ (\text{或}\ \min)\quad z=\sum_{j=1}^{n}\sum_{k=1}^{K_j} f_j(a_{jk})w_{jk}$$

$$\begin{array}{lll}
\text{s.t.} & \displaystyle\sum_{j=1}^{n}\sum_{k=1}^{K_j} g_{jk}(a_{jk})w_{jk}\leqslant b_i, & i=1,2,\cdots,m \\
& 0\leqslant w_{j1}\leqslant y_{j1}, & j=1,2,\cdots,n \\
& 0\leqslant w_{jk}\leqslant y_{j,k-1}+y_{jk}, & k=2,3,\cdots,K_{j-1},\quad j=1,2,\cdots,n \\
& 0\leqslant w_{jK_j}\leqslant y_{j,K_j-1}, & j=1,2,\cdots,n \\
& \displaystyle\sum_{k=1}^{K_j-1} y_{jk}=1, & j=1,2,\cdots,n \\
& \displaystyle\sum_{k=1}^{K_j} w_{jk}=1, & j=1,2,\cdots,n \\
& y_{jk}=(0,1), & k=1,2,\cdots,K_j,\quad j=1,2,\cdots,n
\end{array}$$

该近似问题的变量为 w_{jk} 和 y_{jk}.

上述表达说明了任何可分离问题至少原则上都可以通过混合整数规划来求解. 困难在于随着间断点数目的增加, 约束条件数会迅速增加. 特别是, 因为没有相应可靠的求解大型混合整数规划问题的程序, 因此计算上的可行性值得怀疑.

求解近似模型的另一种方法是采用了**限制基** (restricted basis) 的普通单纯形方法 (第 3 章). 在这种情况下, 去掉了包括 y_{jk} 的额外约束. 限制基修正了单纯形

① 更精确的是用 k_j 代替下标 k 来唯一对应于变量 j, 但为了简化起见我们没有采用这些记号.

方法的最优性条件, 通过选择最好的$(z_{jk}-c_{jk})$ 的进基变量 w_j, 使得两个 w 变量仅当它们相邻时可取正值. 重复这一过程直到最优性条件满足, 或者直到当不违反限制基条件就不可能引入新的 w_{jk} 时为止. 最后一个单纯形表给出了该问题的近似最优解.

混合整数规划方法给出近似问题的全局最优解, 但限制基方法只能保证局部最优解. 此外, 在这两种方法中, 近似解对于原问题来说可能是不可行的. 事实上, 近似模型可能给出不是原问题解空间里的其他点.

例 19.2-1

考虑问题

$$\begin{aligned} \max \quad & z = x_1 + x_2^4 \\ \text{s.t.} \quad & 3x_1 + 2x_2^2 \leqslant 9 \\ & x_1,\ x_2 \geqslant 0 \end{aligned}$$

用 AMPL 或 Excel 规划求解得到的该问题的精确解是 $x_1 = 0$, $x_2 = 2.121\ 32$, $z^* = 20.25$. 为了说明如何利用近似方法, 考虑可分离函数

$$\begin{aligned} f_1(x_1) &= x_1 \\ f_2(x_2) &= x_2^4 \\ g_1(x_1) &= 3x_1 \\ g_2(x_2) &= 2x_2^2 \end{aligned}$$

函数 $f_1(x_1)$ 和 $g_1(x_1)$ 一致, 因为它们已经是线性的了. 在这种情况下, x_1 被看成是所要的变量之一. 考虑 $f_2(x_2)$ 和 $g_2(x_2)$, 我们设置 4 个间断点: $a_{2k} = 0, 1, 2, 3$, 分别对应 $k = 1, 2, 3, 4$. 由于 x_2 的值不能超过 3, 我们得到

k	a_{2k}	$f_2(a_{2k}) = a_{2k}^4$	$g_2(a_{2k}) = 2a_{2k}^2$
1	0	0	0
2	1	1	2
3	2	16	8
4	3	81	18

由此得到

$$\begin{aligned} f_2(x_2) &\approx w_{21}f_2(a_{21}) + w_{22}f_2(a_{22}) + w_{23}f_2(a_{23}) + w_{24}f_2(a_{24}) \\ &\approx 0w_{21} + 1w_{22} + 16w_{23} + 81w_{24} = w_{22} + 16w_{23} + 81w_{24} \end{aligned}$$

类似地,

$$g_2(x_2) \approx 2w_{22} + 8w_{23} + 18w_{24}$$

因此近似问题变成

$$\begin{aligned} \max \quad & z = x_1 + w_{22} + 16w_{23} + 81w_{24} \\ \text{s.t.} \quad & 3x_1 + 2w_{22} + 8w_{23} + 18w_{24} \leqslant 9 \\ & w_{21} + w_{22} + w_{23} + w_{24} = 1 \\ & x_1 \geqslant 0, \ w_{2k} \geqslant 0, \quad k = 1,2,3,4 \end{aligned}$$

w_{2k} 的值, $k = 1,2,3,4$, 必须满足限制基条件.

初始单纯形表如下:

基	x_1	w_{22}	w_{23}	w_{24}	s_1	w_{21}	解
z	-1	-1	-16	-81	0	0	0
s_1	3	2	8	18	1	0	9
w_{21}	0	1	1	1	0	1	1

变量 $s_1(\geqslant 0)$ 是松弛变量 (这个问题恰好有一个显然的初始解. 一般来说, 必须要用人工变量, 见 3.4 节).

根据 z 行系数, w_{24} 是进基变量. 因为 w_{21} 是当前基变量, 而且是正的, 限制基条件要求它必须在 w_{24} 进入解之前离基. 按照可行性条件, s_1 必须是离基变量, 这意味着 w_{24} 不能进入解. 下一个最好的进基变量 w_{23} 要求 w_{21} 离开基本解, 这个条件刚好由可行性条件满足. 因此新表变成

基	x_1	w_{22}	w_{23}	w_{24}	s_1	w_{21}	解
z	-1	15	0	-65	0	16	16
s_1	3	-6	0	10	1	-8	1
w_{23}	0	1	1	1	0	1	1

接下来, w_{24} 是进基变量, 由于 w_{23} 是正的, 因此这是可以的. 单纯形方法表明 s_1 将离基, 因此

基	x_1	w_{22}	w_{23}	w_{24}	s_1	w_{21}	解
z	$\frac{37}{2}$	-24	0	0	$\frac{13}{2}$	-36	$22\frac{1}{2}$
w_{24}	$\frac{3}{10}$	$-\frac{6}{10}$	0	1	$\frac{1}{10}$	$-\frac{8}{10}$	$\frac{1}{10}$
w_{23}	$-\frac{3}{10}$	$\frac{16}{10}$	1	0	$-\frac{1}{10}$	$\frac{18}{10}$	$\frac{9}{10}$

上表说明, w_{21} 和 w_{22} 是候选进基变量. 由于 w_{21} 与基变量 w_{23} 或 w_{24} 不相邻, 因此不能进基. 类似地, 由于 w_{24} 不能离基, 因此 w_{22} 不能进基. 因此最终的单纯形表就是近似问题的最好限制基解.

原问题的最优解是

$$x_1 = 0$$

$$x_2 \approx 2w_{23} + 3w_{24} = 2\left(\frac{9}{10}\right) + 3\left(\frac{1}{10}\right) = 2.1$$

$$z = 0 + 2.1^4 = 19.45$$

值 $x_2 = 2.1$ 近似等于真正的最优值 $(= 2.121\ 32)$.

可分离凸规划 可分离规划的一种特殊情况是, 对所有的 i 和 j, $g_{ij}(x_j)$ 均为凸函数, 这就保证了解空间是凸集. 此外, 若 $f_j(x_j)$ 对所有的 j 是凸的 (求最小值) 或是凹的 (求最大值), 则问题有全局最优解 (见 18.2.2 节的表 18.2). 在这些条件下, 可以采用下面的简化近似方法.

考虑一求最小值问题, 令 $f(x)$ 如图 19.4 所示. 函数 $f_j(x_j)$ 的间断点为 $x_j = a_{jk}, k = 1, 2, \cdots, K_j$. 令 x_{jk} 定义变量 x_j 在区间 $(a_{j,k-1}, a_{jk}), k = 1, 2, \cdots, K_j$ 中的增量, 并令 r_{jk} 为在同样区间里的相应的变化率 (线段的斜率). 则

$$f_j(x_j) \approx \sum_{k=1}^{K_j} r_{jk} x_{jk} + f_j(a_{j0})$$

$$x_j = \sum_{k=1}^{K_j} x_{jk}$$

$$0 \leqslant x_{jk} \leqslant a_{jk} - a_{j,k-1}, k = 1, 2, \cdots, K_j$$

$f_j(x_j)$ 是凸函数的事实保证了 $r_{j1} < r_{j2} < \cdots < r_{jK_j}$. 这意味着在这个求最小值问题中, 对于 $p < q$, 变量 x_{jp} 比 x_{jq} 更有吸引力. 由此, x_{jp} 总是在 x_{jq} 能取得正值之前达到其最大值上限.

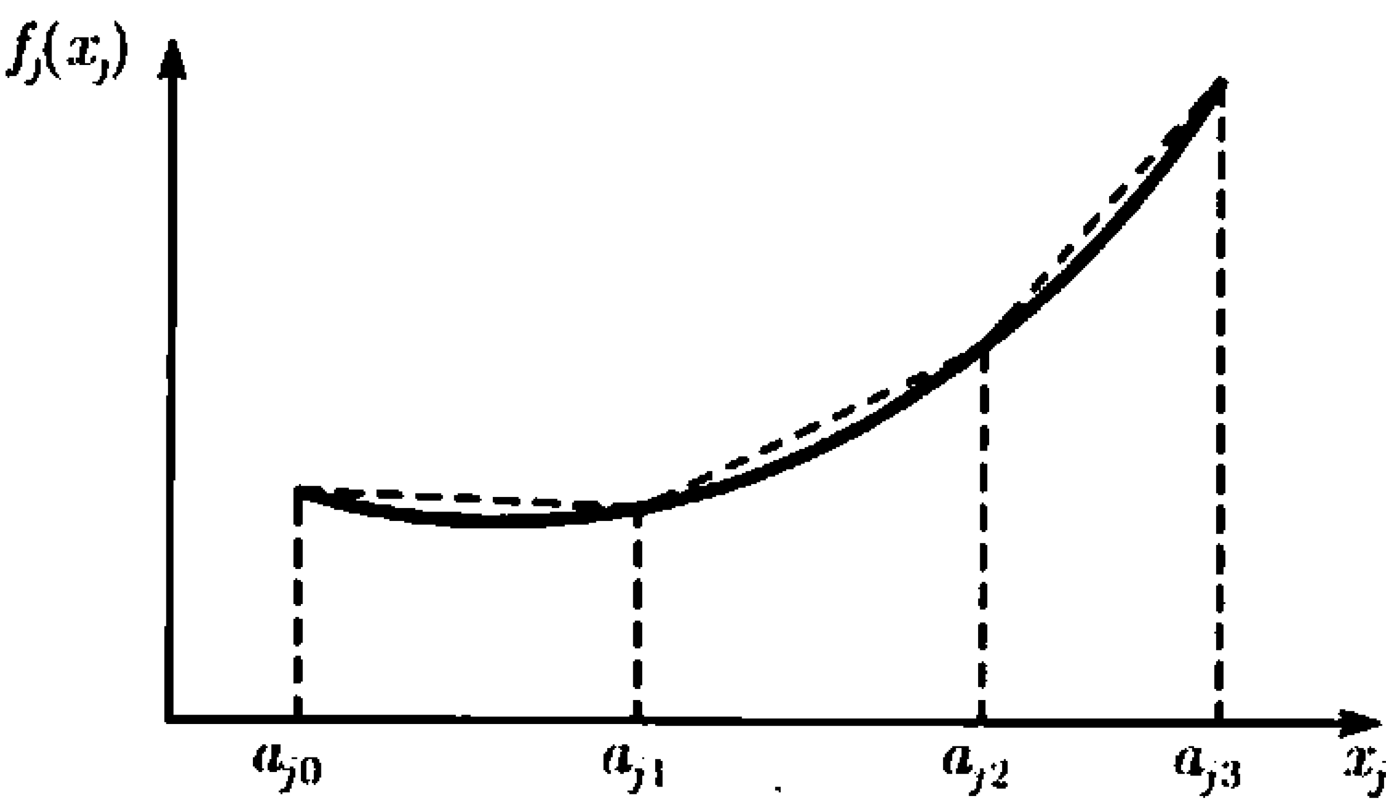


图 19.4 凸函数的分段线性近似

凸约束函数 $g_{ij}(x_j)$ 基本上按同样的方法近似. 令 r_{ijk} 为第 k 个线段对应于 $g_{ij}(x_j)$ 的斜率, 我们得到

$$g_{ij}(x_j) \approx \sum_{k=1}^{K_j} r_{ijk} x_{jk} + g_{ij}(a_{j0})$$

整个问题则表述为

$$\min\ z=\sum_{j=1}^{n}\left(\sum_{k=1}^{K_j} r_{jk}x_{jk}+f_j(a_{j0})\right)$$

$$\text{s.t.}\ \sum_{j=1}^{n}\left(\sum_{k=1}^{K_j} r_{ijk}x_{jk}+g_{ij}(a_{j0})\right)\leqslant b_i,\quad i=1,2,\cdots,m$$

$$0\leqslant x_{jk}\leqslant a_{jk}-a_{j,k-1},\quad k=1,2,\cdots,K_j,\ j=1,2,\cdots,n$$

其中

$$r_{jk}=\frac{f_j(a_{jk})-f_j(a_{j,k-1})}{a_{jk}-a_{j,k-1}},\quad r_{ijk}=\frac{g_{ij}(a_{jk})-g_{ij}(a_{j,k-1})}{a_{jk}-a_{j,k-1}}$$

求最大值问题基本上按同样方法处理. 在这种情况下, $r_{j1}>r_{j2}>\cdots>r_{jK_j}$. 这意味着对于 $p<q$, 变量 x_{jp} 总是在 x_{jq} 能取得正值之前达到其最大值 (证明见习题 19.2A 第 7 题).

这个新问题可用带有上界变量的单纯形方法求解 (见 13.3 节). 限制基的概念就不需要了, 因为函数的凸性 (凹性) 保证了对基变量的正确选择.

例 19.2-2

考虑问题

$$\begin{aligned}\max\quad & z=x_1-x_2\\ \text{s.t.}\quad & 3x_1^4+x_2\leqslant 243\\ & x_1+2x_2^2\leqslant 32\\ & x_1\geqslant 2.1\\ & x_2\geqslant 3.5\end{aligned}$$

这个问题的可分离函数为

$$\begin{aligned} f_1(x_1)&=x_1, & f_2(x_2)&=-x_2\\ g_{11}(x_1)&=3x_1^4, & g_{12}(x_2)&=x_2\\ g_{21}(x_1)&=x_1, & g_{22}(x_2)&=2x_2^2\end{aligned}$$

这些函数满足最小化问题所需的凸性条件. 函数 $f_1(x_1)$, $f_2(x_2)$, $g_{12}(x_2)$, $g_{21}(x_1)$ 已经是线性的, 不需要“做近似”.

变量 x_1 和 x_2(从约束估计的) 的取值范围是 $0\leqslant x_1\leqslant 3$ 和 $0\leqslant x_2\leqslant 4$, 令 $K_1=3$, $K_2=4$, 对应于可分离函数的斜率按下表确定.

对于$j=1$,

k	a_{1k}	$g_{11}(a_{1k})=3a_{1k}^4$	r_{11k}	x_{1k}
0	0	0	—	—
1	1	3	3	x_{11}
2	2	48	45	x_{12}
3	3	243	195	x_{13}

对于$j=2$,

k	a_{2k}	$g_{22}(a_{2k})=2a_{2k}^2$	r_{22k}	x_{2k}
0	0	0	—	—
1	1	2	2	x_{21}
2	2	8	6	x_{22}
3	3	18	10	x_{23}
4	4	32	14	x_{24}

整个问题变成

$$\max\ z=x_1-x_2$$

$$\text{s.t.}\ 3x_{11}+45x_{12}+195x_{13}+x_2 \leqslant 243 \quad (19.1)$$

$$x_1+2x_{21}+6x_{22}+10x_{23}+14x_{24} \leqslant 32 \quad (19.2)$$

$$x_1 \geqslant 2.1 \quad (19.3)$$

$$x_2 \geqslant 3.5 \quad (19.4)$$

$$x_{11}+x_{12}+x_{13}-x_1 = 0 \quad (19.5)$$

$$x_{21}+x_{22}+x_{23}+x_{24}-x_2 = 0 \quad (19.6)$$

$$0 \leqslant x_{1k} \leqslant 1,\quad k=1,2,3 \quad (19.7)$$

$$0 \leqslant x_{2k} \leqslant 1,\quad k=1,2,3,4 \quad (19.8)$$

$$x_1,x_2 \geqslant 0$$

约束 5 和约束 6 用来维持原问题和新问题变量之间的关系. 最优解为

$$z=-0.52, x_1=2.98, x_2=3.5, x_{11}=x_{12}=1, x_{13}=0.98, x_{21}=x_{22}=x_{23}=1, x_{24}=0.5$$

AMPL 程序

AMPL 中非线性问题的建模和线性问题中的完全一样, 但由于非线性函数"捉摸不定"的行为, 使得求解方法大相径庭. 图 19.5 给出了例 19.2-2 原问题的 AMPL 模型 (文件 amplEx19.2-2.txt). 与 LP 问题唯一不同的 (当然除了非线性以外) 是, 你需要指定"合适的"初始变量值, 使得求解迭代收敛. 在图 19.5 中, 任意初始值 $x_1=10$ 和 $x_2=10$ 是由向这两个变量的定义语句中加入`:=10`指定的. 假如你不指

定任何初始值的话, AMPL 程序将不会得到最优解, 并弹出消息框 “too many major iterations (迭代次数过多)”. 尽管在这种情况下给出一个解, 但通常是错误的. 其实, 求解非线性问题最合理的方式是指定不同的初始变量值, 然后确定是否能得出一致的最优解.

```
var x1>=0  :=10;  #inital value = 10
var x2>=0  :=10;  #initial value = 10

maximize z: x1-x2;
subject to c1: 3*x1^4+x2<=243;
subject to c2: x1+2*x2^2<=32;
subject to c3: x1>=2.1;
subject to c4: x2>=3.5;

solve;
display z,x1,x2;
```

图 19.5 例 19.2-2 的 AMPL 模型

评注 AMPL 提供了处理凸可分离规划的特殊程序语句. 若非线性仅出现在目标函数中, 这个程序看起来更加可靠. 否则, 非线性函数的行为会相当复杂, 因此 AMPL 可能会导致约束不可行, 即使事实上并非如此.

习题 19.2A

1. 将下面的问题近似为一个混合整数规划.

$$\begin{aligned} \max\ & z = e^{-x_1} + x_1 + (x_2+1)^2 \\ \text{s.t.}\ & x_1^2 + x_2 \leqslant 3 \\ & x_1, x_2 \geqslant 0 \end{aligned}$$

***2.** 用限制基方法重新求解第 1 题. 并求出最优解.

3. 考虑问题

$$\begin{aligned} \max\ & z = x_1x_2x_3 \\ \text{s.t.}\ & x_1^2 + x_2 + x_3 \leqslant 4 \\ & x_1, x_2, x_3 \geqslant 0 \end{aligned}$$

把该问题近似为一个用限制基方法求解的线性规划.

***4.** 说明下述问题如何变成可分离的.

$$\begin{aligned} \max\ & z = x_1x_2 + x_3 + x_1x_3 \\ \text{s.t.}\ & x_1x_2 + x_2 + x_1x_3 \leqslant 10 \\ & x_1, x_2, x_3 \geqslant 0 \end{aligned}$$

5. 说明下述问题如何变成可分离的.

$$\begin{aligned}\min\ \ & z = e^{2x_1+x_2^2} + (x_3-2)^2 \\ \text{s.t.}\ \ & x_1+x_2+x_3 \leqslant 6 \\ & x_1, x_2, x_3 \geqslant 0\end{aligned}$$

6. 说明下述问题如何变成可分离的.

$$\begin{aligned}\max\ \ & z = e^{x_1x_2} + x_2^2x_3 + x_4 \\ \text{s.t.}\ \ & x_1 + x_2x_3 + x_3 \leqslant 10 \\ & x_1, x_2, x_3 \geqslant 0, x_4\text{无符号限制}\end{aligned}$$

7. 证明在可分离凸规划中, 当 $x_{k-1,j}$ 不在上边界时, 让 $x_{kj} > 0$ 不会是最优的.

8. 作为可分离凸规划问题来求解:

$$\begin{aligned}\min\ \ & z = x_1^4 + x_2 + x_3^2 \\ \text{s.t.}\ \ & x_1^2 + x_2 + x_3^2 \leqslant 4 \\ & |x_1 + x_2| \qquad \leqslant 0 \\ & x_1, x_3 \geqslant 0 \\ & \text{变量 } x_2 \text{ 无符号限制}\end{aligned}$$

9. 作为可分离凸规划问题来求解:

$$\begin{aligned}\min\ \ & z = (x_1-2)^2 + 4(x_2-6)^2 \\ \text{s.t.}\ \ & 6x_1 + 3(x_2+1)^2 \leqslant 12 \\ & x_1, x_2 \geqslant 0\end{aligned}$$

19.2.2 二次规划

定义二次规划模型:

$$\begin{aligned}\max\ \ & z = \boldsymbol{CX} + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{DX} \\ \text{s.t.}\ \ & \boldsymbol{AX} \leqslant \boldsymbol{b},\ \boldsymbol{X} \geqslant \boldsymbol{0}\end{aligned}$$

其中

$$\boldsymbol{X} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}, \quad \boldsymbol{C} = (c_1, c_2, \cdots, c_n), \quad \boldsymbol{b} = (b_1, b_2, \cdots, b_m)^{\mathrm{T}},$$

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn}\end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} d_{11} & \cdots & d_{1n} \\ \vdots & & \vdots \\ d_{n1} & \cdots & d_{nn}\end{pmatrix}$$

函数 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{DX}$ 定义了一个二次型 (见附录 D.3), 矩阵 $\boldsymbol{D}$ 为对称负定矩阵. 这意味着 z 是严格凹的. 约束是线性的, 保证了凸解空间.

求解这个问题要基于 KKT 必要条件. 由于 z 是严格凹的, 且解空间为凸集, 这些条件也是全局解的充分条件 (见 18.2.2 节的表 18.2).

我们将处理这个二次规划问题的极大值情况, 转换成极小化问题是显然的. 该问题可以写成

$$\max \ z = \boldsymbol{CX} + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{DX}$$
$$\text{s.t.} \ \boldsymbol{G}(\boldsymbol{X}) = \begin{pmatrix} \boldsymbol{A} \\ -\boldsymbol{I} \end{pmatrix} \boldsymbol{X} - \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{0} \end{pmatrix} \leqslant \boldsymbol{0}$$

令

$$\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_m)^{\mathrm{T}}, \qquad \boldsymbol{U} = (\mu_1, \mu_2, \cdots, \mu_n)^{\mathrm{T}}$$

为分别对应于约束 $\boldsymbol{AX} - \boldsymbol{b} \leqslant \boldsymbol{0}$ 和 $-\boldsymbol{X} \leqslant \boldsymbol{0}$ 的拉格朗日乘子. 运用 KKT 条件得到

$$\boldsymbol{\lambda} \geqslant \boldsymbol{0}, \ \ \boldsymbol{U} \geqslant \boldsymbol{0}$$
$$\nabla Z - (\boldsymbol{\lambda}^{\mathrm{T}}, \boldsymbol{U}^{\mathrm{T}}) \nabla \boldsymbol{G}(\boldsymbol{X}) = \boldsymbol{0}$$
$$\lambda_i \left(b_i - \sum_{j=1}^{n} a_{ij} x_j \right) = 0, \quad i = 1, 2, \cdots, m$$
$$\mu_j x_j = 0, \quad j = 1, 2, \cdots, n$$
$$\boldsymbol{AX} \leqslant \boldsymbol{b}$$
$$-\boldsymbol{X} \leqslant \boldsymbol{0}$$

现有

$$\nabla z = \boldsymbol{C} + 2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}$$
$$\nabla \boldsymbol{G}(\boldsymbol{X}) = \begin{pmatrix} \boldsymbol{A} \\ -\boldsymbol{I} \end{pmatrix}$$

令 $\boldsymbol{S} = \boldsymbol{b} - \boldsymbol{AX} \geqslant \boldsymbol{0}$ 为约束的松弛变量. 条件简化为

$$-2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D} + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{A} - \boldsymbol{U}^{\mathrm{T}} = \boldsymbol{C}$$
$$\boldsymbol{AX} + \boldsymbol{S} = \boldsymbol{b}$$
$$\mu_j x_j = 0 = \lambda_i S_i, \quad \text{对于所有的 } i, \ j$$
$$\boldsymbol{\lambda}, \ \boldsymbol{U}, \ \boldsymbol{X}, \ \boldsymbol{S} \geqslant \boldsymbol{0}$$

由于 $\boldsymbol{D}^{\mathrm{T}} = \boldsymbol{D}$, 第 1 组方程的转置可写成

$$-2\boldsymbol{DX} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} - \boldsymbol{U} = \boldsymbol{C}^{\mathrm{T}}$$

因此, 必要条件可以合并为

$$\begin{pmatrix} -2\boldsymbol{D} & \boldsymbol{A}^{\mathrm{T}} & -\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{A} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{\lambda} \\ \boldsymbol{U} \\ \boldsymbol{S} \end{pmatrix} = \begin{pmatrix} \boldsymbol{C}^{\mathrm{T}} \\ \boldsymbol{b} \end{pmatrix}$$

$$\mu_j x_j = 0 = \lambda_i S_i, \quad \text{对于所有 } i,\ j$$

$$\boldsymbol{\lambda},\ \boldsymbol{U},\ \boldsymbol{X}, \boldsymbol{S} \geqslant \boldsymbol{0}$$

除条件 $\mu_j x_j = 0 = \lambda_i S_i$ 外, 所有的方程都是 $\boldsymbol{X}$, $\boldsymbol{\lambda}$, $\boldsymbol{U}$, $\boldsymbol{S}$ 的线性函数. 因此, 这个问题等价于解一组带有附加条件 $\mu_j x_j = 0 = \lambda_i S_i$ 的线性方程组. 由于 z 是严格凹的, 且解空间为凸集, 满足所有这些条件的可行解必为唯一的最优解.

这组方程的解可以通过两阶段方法的阶段 I 得到 (见 3.4.2 节). 唯一的约束是满足条件 $\lambda_i S_i = 0 = \mu_j x_j$. 这意味着 λ_i 与 s_i 不能同时为正, μ_j 与 x_j 也不能同时取正值. 这和 19.2.1 节中采用的**限制基** (restricted basis) 方法的思路相同.

若问题有一个可行空间, 阶段 I 将会使得所有人工变量等于 0.

例 19.2-3

考虑问题

$$\begin{aligned} \max\ & z = 4x_1 + 6x_2 - 2x_1^2 - 2x_1x_2 - 2x_2^2 \\ \text{s.t.}\ & x_1 + 2x_2 \leqslant 2 \\ & x_1, x_2 \geqslant 0 \end{aligned}$$

这个问题可写成矩阵形式如下:

$$\begin{aligned} \max\ & z = (4,6)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + (x_1, x_2)\begin{pmatrix} -2 & -1 \\ -1 & -2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ \text{s.t.}\ & (1,2)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \leqslant 2 \\ & x_1, x_2 \geqslant 0 \end{aligned}$$

下面给出 KKT 条件:

$$\begin{pmatrix} 4 & 2 & 1 & -1 & 0 & 0 \\ 2 & 4 & 2 & 0 & -1 & 0 \\ 1 & 2 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \lambda_1 \\ \mu_1 \\ \mu_2 \\ s_1 \end{pmatrix} = \begin{pmatrix} 4 \\ 6 \\ 2 \end{pmatrix}, \quad \mu_1 x_1 = \mu_2 x_2 = \lambda_1 s_1 = 0$$

通过引入人工变量 R_1 和 R_2, 并更新目标行, 得到阶段 1 的初始单纯形表, 因此有

基	x_1	x_2	λ_1	μ_1	μ_2	R_1	R_2	s_1	解
r	6	6	3	-1	-1	0	0	0	10
R_1	4	2	1	-1	0	1	0	0	4
R_2	2	4	2	0	-1	0	1	0	6
s_1	1	2	0	0	0	0	0	1	2

迭代 1 由 $\mu_1 = 0$, 让最有希望的进基变量 x_1 进基, 离基变量为 R_1, 产生下表:

基	x_1	x_2	λ_1	μ_1	μ_2	R_1	R_2	s_1	解
R	0	3	$\frac{3}{2}$	$\frac{1}{2}$	-1	$-\frac{3}{2}$	0	0	4
x_1	1	$\frac{1}{2}$	$\frac{1}{4}$	$-\frac{1}{4}$	0	$\frac{1}{4}$	0	0	1
R_2	0	3	$\frac{3}{2}$	$\frac{1}{2}$	-1	$-\frac{1}{2}$	1	0	4
s_1	0	$\frac{3}{2}$	$-\frac{1}{4}$	$\frac{1}{4}$	0	$-\frac{1}{4}$	0	1	1

迭代 2 由 $\mu_2 = 0$, 让最有希望的进基变量 x_2 进基, 产生下表:

基	x_1	x_2	λ_1	μ_1	μ_2	R_1	R_2	s_1	解
r	0	0	2	0	-1	-1	0	-2	2
x_1	1	0	$\frac{1}{3}$	$-\frac{1}{3}$	0	$\frac{1}{3}$	0	$-\frac{1}{3}$	$\frac{2}{3}$
R_2	0	0	2	0	-1	0	1	-2	2
x_2	0	1	$-\frac{1}{6}$	$\frac{1}{6}$	0	$-\frac{1}{6}$	0	$\frac{2}{3}$	$\frac{2}{3}$

迭代 3 由 $s_1 = 0$, λ_1 可进入解, 得到:

基	x_1	x_2	λ_1	μ_1	μ_2	R_1	R_2	s_1	解
r	0	0	0	0	0	-1	-1	0	0
x_1	1	0	0	$-\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{3}$	$-\frac{1}{6}$	0	$\frac{1}{3}$
λ_1	0	0	1	0	$-\frac{1}{2}$	0	$\frac{1}{2}$	-1	1
x_2	0	1	0	$\frac{1}{6}$	$-\frac{1}{12}$	$-\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{2}$	$\frac{5}{6}$

最后一个表给出了阶段 1 的最优解. 因为 $r = 0$, 解 $x_1^* = \frac{1}{3}$, $x_2^* = \frac{5}{6}$ 是可行的. 从原问题计算出的 z 的最优值为 4.16.

Excel 规划求解程序

图 19.6 显示用 Excel 规划求解计算的例 19.2-3 的解 (文件 excelQP.xls). 输入数据的方式与线性规划类似 (见 2.4.2 节). 主要的区别是输入非线性函数的方式. 具体说来, 在例 19.2-3 中, 非线性目标函数

$$z = 4x_1 + 6x_2 - 2x_1^2 - 2x_1x_2 - 2x_2^2$$

在目标单元格 D5 中输入为

```
=4*B10+6*c10-2*B10^2-2*B10*C10-2*C10^2
```

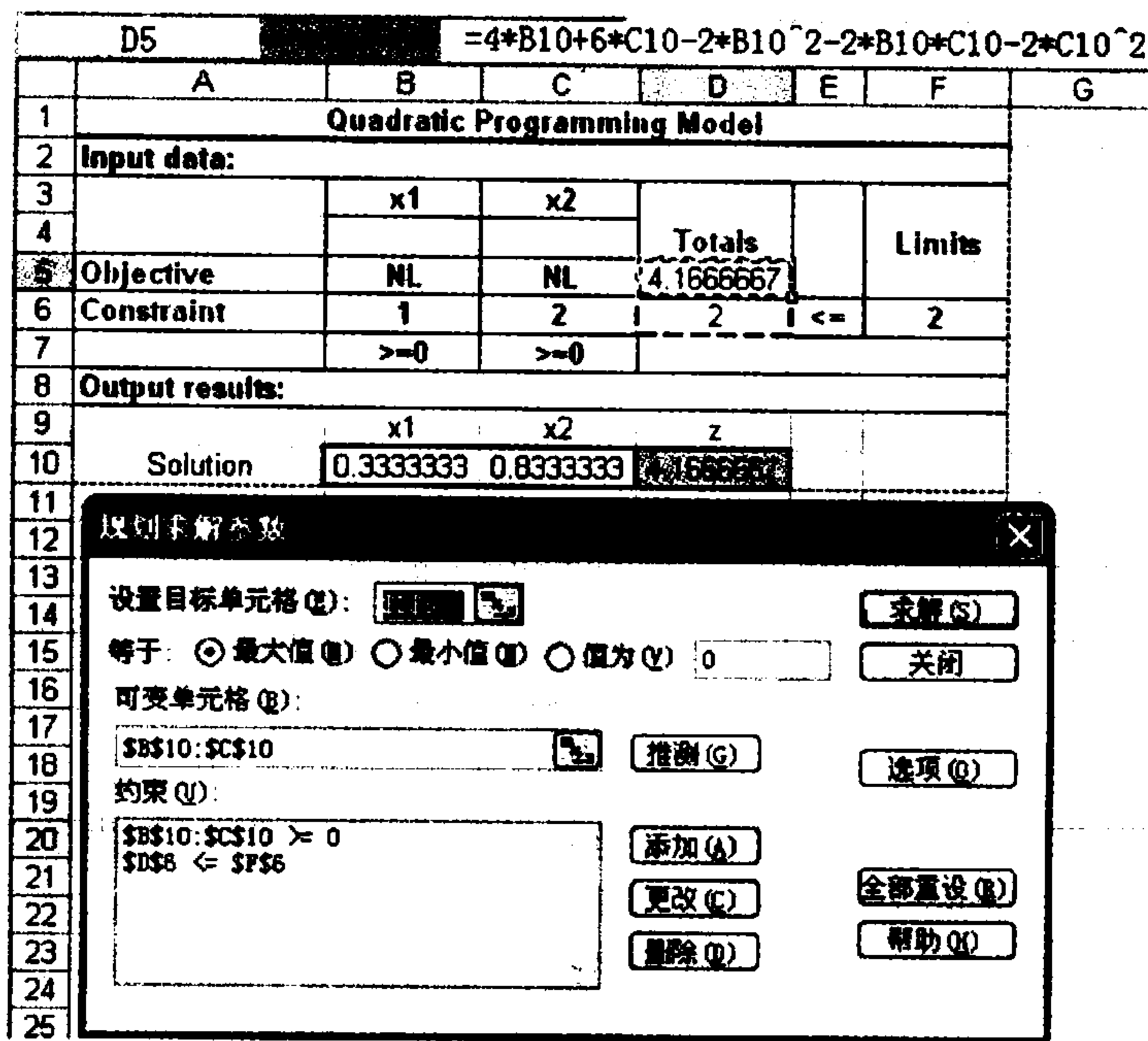


图 19.6 例 19.2-3 二次规划的 Excel 规划求解

因此, 变化单元格为 B10:C10[$\equiv (x_1, x_2)$]. 注意单元格 B5:C5 在模型中根本没有用到. 为了更清楚, 我们输入 NL 来指示相应的约束是非线性的. 另外, 可以在选项对话框中指定变量的非负性, 也可以直接加入非负性约束.

习题 19.2B

***1. 考虑问题:**

$$
\begin{aligned}
\max \quad & z = 6x_1 + 3x_2 - 4x_1x_2 - 2x_1^2 - 3x_2^2 \\
\text{s.t.} \quad & x_1 + x_2 \leqslant 1 \\
& 2x_1 + 3x_2 \leqslant 4 \\
& x_1, x_2 \geqslant 0
\end{aligned}
$$

说明 z 是严格凹函数, 并用二次规划算法求解该问题.

***2. 考虑问题:**

$$
\begin{aligned}
\min \quad & z = 2x_1^2 + 2x_2^2 + 3x_3^2 + 2x_1x_2 + 2x_2x_3 + x_1 - 3x_2 - 5x_3 \\
\text{s.t.} \quad & x_1 + x_2 + x_3 \geqslant 1 \\
& 3x_1 + 2x_2 + x_3 \leqslant 6 \\
& x_1, x_2, x_3 \geqslant 0
\end{aligned}
$$

说明 z 是严格凸函数, 并用二次规划算法求解该问题.

19.2.3 机会约束规划

在机会约束规划情形下, 约束的系数为随机变量, 而且约束是由最小概率来实现的. 数学上, 这一问题定义为

$$\max \ z = \sum_{j=1}^{n} c_j x_j$$

$$\text{s.t. } \mathrm{P}\left\{\sum_{j=1}^{n} a_{ij} x_j \leqslant b_i\right\} \geqslant 1-\alpha_i, \quad i=1,2,\cdots,m, \ x_j \geqslant 0, \text{ 对于所有的 } j$$

参数 a_{ij} 和 b_i 都是随机变量, 约束 i 用不超过 $1-\alpha_i$, $0<\alpha_i<1$ 的最小概率实现.

考虑 3 种情况:

(1) 只有 a_{ij} 是随机的, 对于所有的 i, j;

(2) 只有 b_i 是随机的, 对于所有的 i;

(3) a_{ij} 和 b_i 都是随机的, 对于所有的 i, j.

在所有这 3 种情况中, 假定这些参数服从已知平均值和方差的正态分布.

情况 1 每个 a_{ij} 都是正态分布的, 平均值为 $\mathrm{E}\{a_{ij}\}$, 方差 $\mathrm{var}\{a_{ij}\}$, a_{ij} 和 $a_{i'j'}$ 的协方差为 $\mathrm{cov}\{a_{ij}, a_{i',j'}\}$. 考虑

$$\mathrm{P}\left\{\sum_{j=1}^{n} a_{ij} x_j \leqslant b_i\right\} \geqslant 1-\alpha_i$$

并定义

$$h_i = \sum_{j=1}^{n} a_{ij} x_j$$

则 h_i 服从正态分布, 且

$$\mathrm{E}\{h_i\} = \sum_{j=1}^{n} \mathrm{E}\{a_{ij}\} x_j, \qquad \mathrm{var}\{h_i\} = \boldsymbol{X}^{\mathrm{T}} \boldsymbol{D}_i \boldsymbol{X}$$

其中

$$\boldsymbol{X} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$$

$$\boldsymbol{D}_i = \text{第 } i \text{ 个协方差矩阵} = \begin{pmatrix} \mathrm{var}\{a_{i1}\} & \cdots & \mathrm{cov}\{a_{i1}, a_{in}\} \\ \vdots & & \vdots \\ \mathrm{cov}\{a_{in}, a_{i1}\} & \cdots & \mathrm{var}\{a_{in}\} \end{pmatrix}$$

由此得

$$\mathrm{P}\{h_i \leqslant b_i\} = \mathrm{P}\left\{\frac{h_i - \mathrm{E}\{h_i\}}{\sqrt{\mathrm{var}\{h_i\}}} \leqslant \frac{b_i - \mathrm{E}\{h_i\}}{\sqrt{\mathrm{var}\{h_i\}}}\right\} \geqslant 1-\alpha_i$$

其中 $\frac{h_i-\mathrm{E}\{h_i\}}{\sqrt{\mathrm{var}\{h_i\}}}$ 服从均值为 0、方差为 1 的标准正态分布. 这意味着

$$\mathrm{P}\{h_i \leqslant b_i\} = F\left(\frac{b_i - \mathrm{E}\{h_i\}}{\sqrt{\mathrm{var}\{h_i\}}}\right)$$

其中 F 表示标准正态分布的 CDF.

令 K_{α_i} 为标准正态值, 使得

$$F(K_{\alpha_i}) = 1 - \alpha_i$$

则若使 $\mathrm{P}\{h_i \leqslant b_i\} \geqslant 1 - \alpha_i$ 成立, 当且仅当

$$\frac{b_i - \mathrm{E}\{h_i\}}{\sqrt{\mathrm{var}\{h_i\}}} \geqslant K_{\alpha_i}$$

由此得到下面的确定性约束:

$$\sum_{j=1}^{n} \mathrm{E}\{a_{ij}\}x_j + K_{\alpha_i}\sqrt{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}_i\boldsymbol{X}} \leqslant b_i$$

对于系数 a_{ij} 都是独立的特殊情况,

$$\mathrm{cov}\{a_{ij}, a_{i'j'}\} = 0$$

且最后一个约束简化成

$$\sum_{j=1}^{n} \mathrm{E}\{a_{ij}\}x_j + K_{\alpha_i}\sqrt{\sum_{j=1}^{n} \mathrm{var}\{a_{ij}\}x_j^2} \leqslant b_i$$

利用代换

$$y_i = \sqrt{\sum_{j=1}^{n} \mathrm{var}\{a_{ij}\}x_j^2}, \quad \text{对于所有的 } i$$

可以把这个约束写成可分离规划形式 (19.2.1 节). 因此, 原约束等价于

$$\sum_{j=1}^{n} \mathrm{E}\{a_{ij}\}x_j + K_{\alpha_i}y_i \leqslant b_i$$

和

$$\sum_{j=1}^{n} \mathrm{var}\{a_{ij}\}x_j^2 - y_i^2 = 0$$

情况 2 只有 b_i 是正态分布的, 平均值为 $\mathrm{E}\{b_i\}$, 方差 $\mathrm{var}\{b_i\}$. 和情况 1 的分析类似, 考虑随机约束

$$\mathrm{P}\left\{b_i \geqslant \sum_{j=1}^{n} a_{ij}x_j\right\} \geqslant \alpha_i$$

和情况 1 一样,

$$\mathrm{P}\left\{\frac{b_i - \mathrm{E}\{b_i\}}{\sqrt{\mathrm{var}\{b_i\}}} \geqslant \frac{\sum\limits_{j=1}^{n} a_{ij}x_j - \mathrm{E}\{b_i\}}{\sqrt{\mathrm{var}\{b_i\}}}\right\} \geqslant \alpha_i$$

上式成立仅当

$$\frac{\sum\limits_{j=1}^{n} a_{ij}x_j - \mathrm{E}\{b_i\}}{\sqrt{\mathrm{var}\{b_i\}}} \leqslant K_{\alpha_i}$$

因此, 这个随机约束等价于确定性的线性约束

$$\sum_{j=1}^{n} a_{ij}x_j \leqslant \mathrm{E}\{b_i\} + K_{\alpha_i}\sqrt{\mathrm{var}\{b_i\}}$$

情况 3 在这种情况下, 所有的 a_{ij} 和 b_i 都是正态随机变量. 考虑约束

$$\sum_{j=1}^{n} a_{ij}x_j \leqslant b_i$$

上式可写成

$$\sum_{j=1}^{n} a_{ij}x_j - b_i \leqslant 0$$

由于所有的 a_{ij} 和 b_i 都是正态的, $\sum\limits_{j=1}^{n} a_{ij}x_j - b_i$ 也是正态的. 这表明机会约束化简为情况 1 的情形, 可以按照类似的方法处理.

例 19.2-4

考虑机会约束问题

$$\begin{aligned}
\max\ & z = 5x_1 + 6x_2 + 3x_3 \\
\text{s.t.}\ & \mathrm{P}\{a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \leqslant 8\} \geqslant 0.95 \\
& \mathrm{P}\{5x_1 + x_2 + 6x_3 \leqslant b_2\} \geqslant 0.10 \\
& x_1, x_2, x_3 \geqslant 0
\end{aligned}$$

设参数 a_{1j}, $j = 1, 2, 3$ 为独立正态分布的随机变量, 其平均值和方差分别是:

$$\mathrm{E}\{a_{11}\} = 1,\quad \mathrm{E}\{a_{12}\} = 3,\quad \mathrm{E}\{a_{13}\} = 9$$

$$\text{var}\{a_{11}\} = 25, \quad \text{var}\{a_{12}\} = 16, \quad \text{var}\{a_{13}\} = 4$$

参数 b_2 服从正态分布, 平均值为 7, 方差为 9.

根据附录 B 中的标准正态表 (或 excelStatTables.xls),

$$K_{\alpha_1} = K_{0.05} \approx 1.645, K_{\alpha_2} = K_{0.10} \approx 1.285$$

对于第 1 个约束, 等价的确定性约束为

$$x_1 + 3x_2 + 9x_3 + 1.645\sqrt{25x_1^2 + 16x_2^2 + 4x_3^2} \leqslant 8$$

对于第 2 个约束

$$5x_1 + x_2 + 6x_3 \leqslant 7 + 1.285(3) = 10.855$$

上述问题可作为非线性规划用 AMPL 或 Excel 规划求解来求解, 也可以转化为下面的可分离规划:

$$y^2 = 25x_1^2 + 16x_2^2 + 4x_3^2$$

问题变成

$$\begin{aligned}
\max \quad & z = 5x_1 + 6x_2 + 3x_3 \\
\text{s.t.} \quad & x_1 + 3x_2 + 9x_3 + 1.645y \leqslant 8 \\
& 25x_1^2 + 16x_2^2 + 4x_3^2 - y^2 = 0 \\
& 5x_1 + x_2 + 6x_3 \leqslant 10.855 \\
& x_1, x_2, x_3, y \geqslant 0
\end{aligned}$$

这个问题可以用可分离规划来求解.

Excel 规划求解程序

例 19.2-4 非线性问题的 Excel 最优解如图 19.7 所示 (文件 excelCCP.xls). 只有约束的左端项是非线性的, 在单元 F7 中输入

```
=25*B12^2+16*c12^2+4*D12^2-E12^2
```

习题 19.2C

*1. 把下面的随机问题转化为一个等价的确定性模型.

$$\begin{aligned}
\max \quad & z = x_1 + 2x_2 + 5x_3 \\
\text{s.t.} \quad & \text{P}\{a_1x_1 + 3x_2 + a_3x_3 \leqslant 10\} \geqslant 0.9 \\
& \text{P}\{7x_1 + 5x_2 + x_3 \leqslant b_2\} \geqslant 0.1 \\
& x_1, x_2, x_3 \geqslant 0
\end{aligned}$$

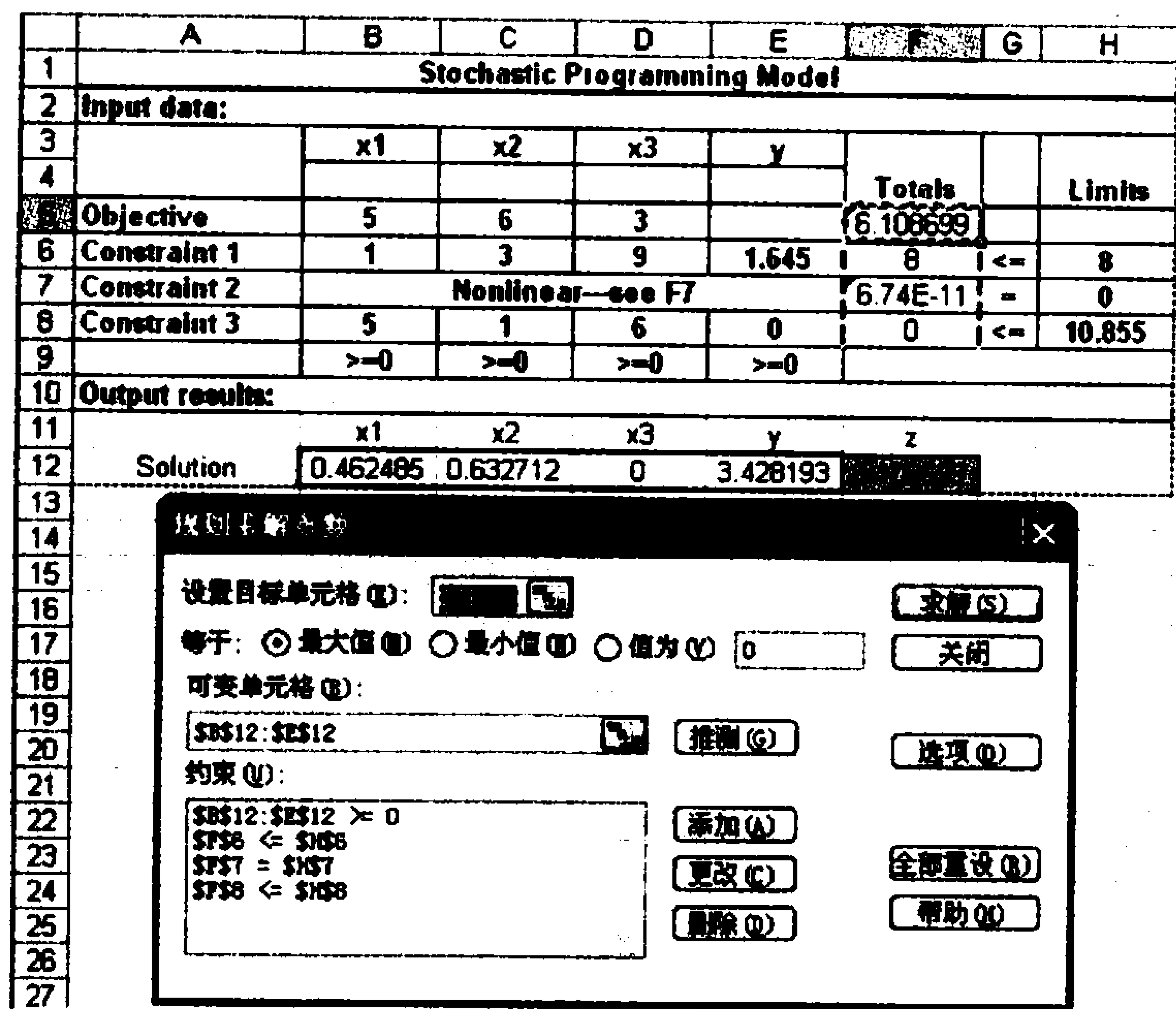


图 19.7　例 19.2-4 机会约束规划的 Excel 规划求解

设 a_1 和 a_3 是独立正态分布的随机变量, 平均值 $\mathrm{E}\{a_1\}=2$, $\mathrm{E}\{a_3\}=5$, 方差 $\mathrm{var}\{a_1\}=9$, $\mathrm{var}\{a_3\}=16$. b_2 正态分布, 平均值为 15, 方差为 25.

2. 考虑以下随机规划模型:

$$
\begin{aligned}
\max\ \ & z=x_1+x_2^2+x_3\\
\text{s.t.}\ \ & \mathrm{P}\{x_1^2+a_2x_2^3+a_3\sqrt{x_3}\leqslant 10\}\geqslant 0.9\\
& x_1,x_2,x_3\geqslant 0
\end{aligned}
$$

参数 a_2 和 a_3 是独立正态分布的随机变量, 平均值分别为 5 和 2, 方差分别为 16 和 25. 将该问题转化为一种 (确定性的) 可分离规划形式.

19.2.4　线性组合方法

本节求解下述问题, 其中所有约束为线性:

$$
\begin{aligned}
\max\ \ & z=f(\boldsymbol{X})\\
\text{s.t.}\ \ & \boldsymbol{AX}\leqslant \boldsymbol{b},\quad \boldsymbol{X}\geqslant \boldsymbol{0}
\end{aligned}
$$

求解方法基于最速上升 (梯度) 方法 (19.1.2 节). 但是, 由梯度向量指定的方向可能不会产生该约束问题的一个可行解. 另外, 在最优 (约束) 点处的梯度向量不一定为零. 因此必须对最速上升方法进行修改, 以便处理带有约束的情况.

令 $\boldsymbol{X}_k$ 为第 k 步迭代可行的试验点, 目标函数 $f(\boldsymbol{X})$ 可在 $\boldsymbol{X}_k$ 邻域内用泰勒

级数展开, 得到

$$f(\boldsymbol{X}) \approx f(\boldsymbol{X}_k) + \nabla f(\boldsymbol{X}_k)(\boldsymbol{X} - \boldsymbol{X}_k) = (f(\boldsymbol{X}_k) - \nabla f(\boldsymbol{X}_k)\boldsymbol{X}_k) + \nabla f(\boldsymbol{X}_k)\boldsymbol{X}$$

求解方法要求确定一个可行点 $\boldsymbol{X} = \boldsymbol{X}^*$, 使得在问题的 (线性) 约束下求出 $f(\boldsymbol{X})$ 的最大值. 由于 $f(\boldsymbol{X}_k) - \nabla f(\boldsymbol{X}_k)\boldsymbol{X}_k$ 是一个常数, 因此求 $\boldsymbol{X}^*$ 的问题简化为求解下述线性规划:

$$\max \ w_k(\boldsymbol{X}) = \nabla f(\boldsymbol{X}_k)\boldsymbol{X}$$
$$\text{s.t.} \ \boldsymbol{AX} \leqslant \boldsymbol{b}, \quad \boldsymbol{X} \geqslant \boldsymbol{0}$$

因为 w_k 是由在 $\boldsymbol{X}_k$ 点处 $f(\boldsymbol{X})$ 的梯度构造成的, 所以当且仅当 $w_k(\boldsymbol{X}^*) > w_k(\boldsymbol{X}_k)$ 时, 才能保证得到改善的解点. 根据泰勒展开式, 只有当 $\boldsymbol{X}^*$ 位于 $\boldsymbol{X}_k$ 的邻域内, 上述条件才能保证 $f(\boldsymbol{X}^*) > f(\boldsymbol{X}_k)$. 但是由于 $w_k(\boldsymbol{X}^*) > w_k(\boldsymbol{X}_k)$, 故必存在线段 $(\boldsymbol{X}_k, \boldsymbol{X}^*)$ 上的一点 $\boldsymbol{X}_{k+1}$, 使得 $f(\boldsymbol{X}_{k+1}) > f(\boldsymbol{X}_k)$. 我们的目的是求出 $\boldsymbol{X}_{k+1}$. 定义

$$\boldsymbol{X}_{k+1} = (1-r)\boldsymbol{X}_k + r\boldsymbol{X}^* = \boldsymbol{X}_k + r(\boldsymbol{X}^* - \boldsymbol{X}_k), \ 0 < r \leqslant 1$$

这就意味着, $\boldsymbol{X}_{k+1}$ 是 $\boldsymbol{X}_k$ 与 $\boldsymbol{X}^*$ 的线性组合. 因为 $\boldsymbol{X}_k$ 和 $\boldsymbol{X}^*$ 是一个凸解空间中的两个可行点, 因此 $\boldsymbol{X}_{k+1}$ 也是可行点. 按照最速上升方法 (19.1.2 节), 参数 r 代表步长.

确定 $\boldsymbol{X}_{k+1}$ 是要让 $f(\boldsymbol{X})$ 达到最大. 由于 $\boldsymbol{X}_{k+1}$ 只是 r 的函数, 确定 $\boldsymbol{X}_{k+1}$ 就是求极大值

$$h(r) = f(\boldsymbol{X}_k + r(\boldsymbol{X}^* - \boldsymbol{X}_k))$$

重复上述过程, 直到第 k 次迭代, 使得 $w_k(\boldsymbol{X}^*) \leqslant w_k(\boldsymbol{X}_k)$. 在该点处, 不可能再有改善, 过程终止, $\boldsymbol{X}_k$ 为最好的解点.

连续迭代中生成的线性规划问题仅仅在目标函数的系数上有差别, 因此 4.5 节中介绍的后最优分析方法可用来进行有效的计算.

例 19.2-5

考虑例 19.2-3 的二次规划问题.

$$\max \ f(\boldsymbol{X}) = 4x_1 + 6x_2 - 2x_1^2 - 2x_1x_2 - 2x_2^2$$
$$\text{s.t.} \ x_1 + 2x_2 \leqslant 2$$
$$x_1, x_2 \geqslant 0$$

令初始试验点 $\boldsymbol{X}_0 = (\frac{1}{2}, \frac{1}{2})$, 它是可行的. 现有

$$\nabla f(\boldsymbol{X}) = (4 - 4x_1 - 2x_2, 6 - 2x_1 - 4x_2)$$

迭代 1

$$\nabla f(\boldsymbol{X}_0) = (1,3)$$

相应的线性规划是在原问题的约束下求 $w_1 = x_1 + 3x_2$ 的最大值. 这给出最优解 $\boldsymbol{X}^* = (0,1)$. w_1 在 $\boldsymbol{X}_0$ 和 $\boldsymbol{X}^*$ 点的值分别为 2 和 3. 因此新的试验点确定为

$$\boldsymbol{X}_1 = \left(\frac{1}{2},\frac{1}{2}\right) + r\left[(0,1) - \left(\frac{1}{2},\frac{1}{2}\right)\right] = \left(\frac{1-r}{2},\frac{1+r}{2}\right)$$

因此

$$h(r) = f\left(\frac{1-r}{2},\frac{1+r}{2}\right)$$

的最大值点为 $r_1 = 1$. 因此有 $\boldsymbol{X}_1 = (0,1)$, $f(\boldsymbol{X}_1) = 4$.

迭代 2

$$\nabla f(\boldsymbol{X}_1) = (2,2)$$

新的线性规划问题的目标函数为 $w_2 = 2x_1 + 2x_2$. 这个问题的最优解为 $\boldsymbol{X}^* = (2,0)$. 由于 w_2 在 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}^*$ 点的值分别为 2 和 4, 因此新的试验点确定为

$$\boldsymbol{X}_2 = (0,1) + r[(2,0) - (0,1)] = (2r, 1-r)$$

从而

$$h(r) = f(2r, 1-r)$$

的最大值点为 $r_2 = \frac{1}{6}$. 因此有 $\boldsymbol{X}_2 = \left(\frac{1}{3},\frac{5}{6}\right)$, $f(\boldsymbol{X}_2) \approx 4.16$.

迭代 3

$$\nabla f(\boldsymbol{X}_2) = (1,2)$$

相应的目标函数为 $w_3 = x_1 + 2x_2$. 这个问题的最优解为 $\boldsymbol{X}^* = (0,1)$ 或 $\boldsymbol{X}^* = (2,0)$. 由于 w_3 在这两个点的值都等于 $\boldsymbol{X}_2$ 的值, 因此不可能有进一步的改善. 近似最优解为 $\boldsymbol{X}_2 = \left(\frac{1}{3},\frac{5}{6}\right)$, $f(\boldsymbol{X}_2) \approx 4.16$, 这恰好就是精确最优解.

习题 19.2D

1. 用线性组合方法求解以下问题.

$$\begin{aligned} \min\quad & f(\boldsymbol{X}) = x_1^3 + x_2^3 - 3x_1x_2 \\ \text{s.t.}\quad & 3x_1 + x_2 \leqslant 3 \\ & 5x_1 - 3x_2 \leqslant 5 \\ & x_1, x_2 \geqslant 0 \end{aligned}$$

19.2.5 SUMT 算法

本节介绍一种更一般性的梯度方法, 假定目标函数 $f(\boldsymbol{X})$ 是凹的, 每个约束函数 $g_i(\boldsymbol{X})$ 都是凸函数. 此外, 解空间必须有内点. 这也就排除了等式约束的情形.

SUMT 算法 (Sequential Unconstrained Maximization Technique, 序贯无约束最大化算法) 是基于把约束问题变换成等价的无约束问题, 多少有些类似于拉格朗日乘子方法, 经过变换后的问题就可以用最速上升方法求解了 (19.1.2 节).

为了更进一步说明这一概念, 考虑新问题

$$p(\boldsymbol{X}, t) = f(\boldsymbol{X}) + t\left(\sum_{i=1}^{m} \frac{1}{g_i(\boldsymbol{X})} - \sum_{j=1}^{n} \frac{1}{x_j}\right)$$

其中 t 是一个非负参数. 第 2 个求和符号表示非负约束, 必须写成 $-x_j \leqslant 0$ 的形式, 以便和原来的约束一致. 由于 $g_i(\boldsymbol{X})$ 是凸的, 则 $\frac{1}{g_i(\boldsymbol{X})}$ 是凹的, 这意味着 $p(\boldsymbol{X}, t)$ 是 $\boldsymbol{X}$ 的凹函数. 这样, $p(\boldsymbol{X}, t)$ 具有唯一的最优解. 求原约束问题的最优解等价于求 $p(\boldsymbol{X}, t)$ 的最优解.

本算法从任意选择 t 的非负初始值开始, 选择初始点 $\boldsymbol{X}_0$ 作为第一个试点. 这个点必为一内点, 即必须不在解空间的边界上. 给定 t 的值, 用最速上升方法求 $p(\boldsymbol{X}, t)$ 相应的最优解 (最大值点).

这个新的解点将永远保持内点, 因为假如解点接近于边界的话, 则至少有一个函数 $\frac{1}{g_i(\boldsymbol{X})}$ 或 $-\frac{1}{x_j}$ 将会取得非常大的负值. 因为我们的目标是求 $p(\boldsymbol{X}, t)$ 的最大值, 这样的解点自动被排除掉. 其主要的结果是, 连续的解点将总是内点, 这样, 这一问题可以总是作为无约束问题的情况来处理.

一旦得到了与给定 t 值对应的最优解, 就生成一个新的 t 值, (用最速上升法) 重复这一最优化过程. 若 t' 是 t 的当前值, 则必须选择下一个值 t'' 使得 $0 < t'' < t'$.

SUMT 算法的结束条件是, 对于两个连续的 t 值, 由求 $p(\boldsymbol{X}, t)$ 的最大值而得到的 $\boldsymbol{X}$ 的最优值近似相等. 在这一点处, 进一步的试点几乎不产生改善.

SUMT 算法的实际运用还涉及更多的细节, 特别是, t 的初始值的选择是影响收敛速度的一个重要因素. 进一步地, 确定初始内点也需要特殊的技巧. 这些细节可以参考 Fiacco and McCormick(1968).

参考文献

Bazaraa, M., H. Sherall, and C. Shetty, *Nonlinear Programming, Theory and Algorithms,* 2nd ed., Wiley, New York, 1993.

Beightler, C., D. Phillips, and D. Wilde, *Foundations of Optimization,* 2nd ed., Prentice Hall, Upper Saddle River, NJ, 1979.

Fiacco, A., and G. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Wiley, New York, 1968.

Luenberger, D., *Linear and Nonlinear Programming*, Kluwer Academic Publishers, Boston, 2003.

Rardin, D., *Optimization in Operations Research*, Prentice Hall, Upper Saddle River, NJ, 1998.

第 20 章　网络与线性规划算法进阶

本章导读　本章分为 3 节. 20.1 节主要介绍已经在第 6 章中出现过的带有容量限制的最小费用流问题. 这种网络流问题涉及各条弧上有容量的限制, 同时在不同的节点上也可以有外部的流进入或者内部的流离开, 目标是找到满足容量限制和额外流约束条件下的最小费用流. 第 6 章中的最短路径问题和最大流问题都是它的特例. 本章其余的两节讨论的主要是第 13 章提到的问题, 求解大规模的 LP 问题. 源于 20 世纪 60 年代的分解算法, 是将有特殊结构的大规模 LP 问题分解为易于计算的子问题. 由于这种算法会带来大量的计算, 所以它只是在理论上比较经典, 而实际上却很难进行计算. 另一方面, 20 世纪 80 年代 Karmarkar 提出的内点算法却是理论和计算都可行的. 与单纯形算法不同的是, 内点算法的迭代是通过对可行解区域进行分割得到的. 从计算复杂性的角度来看, 单纯形算法是指数时间的算法, 而内点算法则是多项式时间的算法. 因此, 内点算法更适用于大规模的线性规划.

本章包括 7 个例题、34 个节后习题、2 个 AMPL 模型、1 个 Excel 规划求解模型和 2 个案例, 其中案例在本书的附录 E 中. AMPL/Excel/Solver/TORA 程序放在文件夹 ch20Files 中.

20.1　带有容量限制的最小费用流问题

研究带有容量限制的最小费用流问题基于下面的假设.

(1) 所有的弧都是有向的 (单向的).

(2) 每条弧上有一个相应的 (非负的) 单位流费用.

(3) 弧上也可以有容量限制的正的下界约束.

(4) 网络中的任何一个节点都可以作为源点, 也可以作为汇点.

新的模型就是要寻找满足弧上容量限制和节点上供应量和需求量条件的最小费用流. 首先建立带有容量限制的网络流模型以及它的等价的线性规划表示, 然后根据这个线性规划表示来设计用于求解网络流模型的一种特殊的限制单纯形算法.

20.1.1　网络表示

考虑一个带有容量限制的网络 $G=(N,A)$, 其中 N 是节点的集合, A 是弧的集合. 定义

$$x_{ij} = \text{从节点 } i \text{ 到节点 } j \text{ 的总流量}$$
$$u_{ij}(l_{ij}) = \text{弧 } (i,j) \text{ 上容量的上界 (下界)}$$
$$c_{ij} = \text{从节点 } i \text{ 到节点 } j \text{ 的单位流费用}$$
$$f_i = \text{节点 } i \text{ 上的网络流}$$

图 20.1 直观地描述了在弧 (i,j) 上定义的各个量. 标号 $[f_i]$ 的意义是, 当节点 i 有网络供应量时, $[f_i]$ 取正值; 当节点 i 有网络需求量时, $[f_i]$ 取负值.

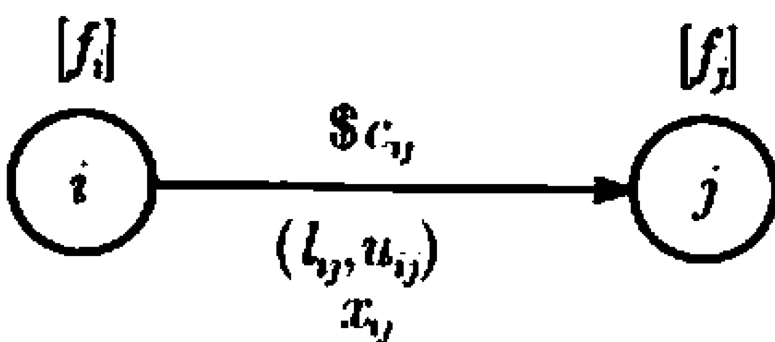


图 20.1　带有外部流的容量限制弧

例 20.1-1

GrainCo 从 3 个仓库为 3 个家禽农场提供饲料. 3 个仓库的供应量分别是 100, 200, 50 千蒲式耳 (1 蒲式耳等于 8 加仑), 3 个农场的需求量分别是 150, 80, 120 千蒲式耳. GrainCo 主要是通过铁路将饲料运送到农场, 另外部分路段也可以使用卡车运输.

图 20.2 给出了在仓库和农场之间可用的路线. 仓库用节点 1, 2, 3 表示, 它们的供应量分别为 [100], [200], [50]; 农场用节点 4, 5, 6 表示, 需求量分别是 [−150], [−80], [−120]. 同时我们允许在仓库之间有几条路线, 如图所示. 弧 (1, 4), (3, 4), (5, 6) 是卡车路线, 这些路线上有容量的上界和下界限制. 例如, 路线 (1, 4) 上的容量限制是介于 50 和 80 千蒲式耳之间. 其他所有的路线都是铁路, 相应的最大容量实际上是没有限制的. 图中同时给出了每条弧上相应的单位运输费用.

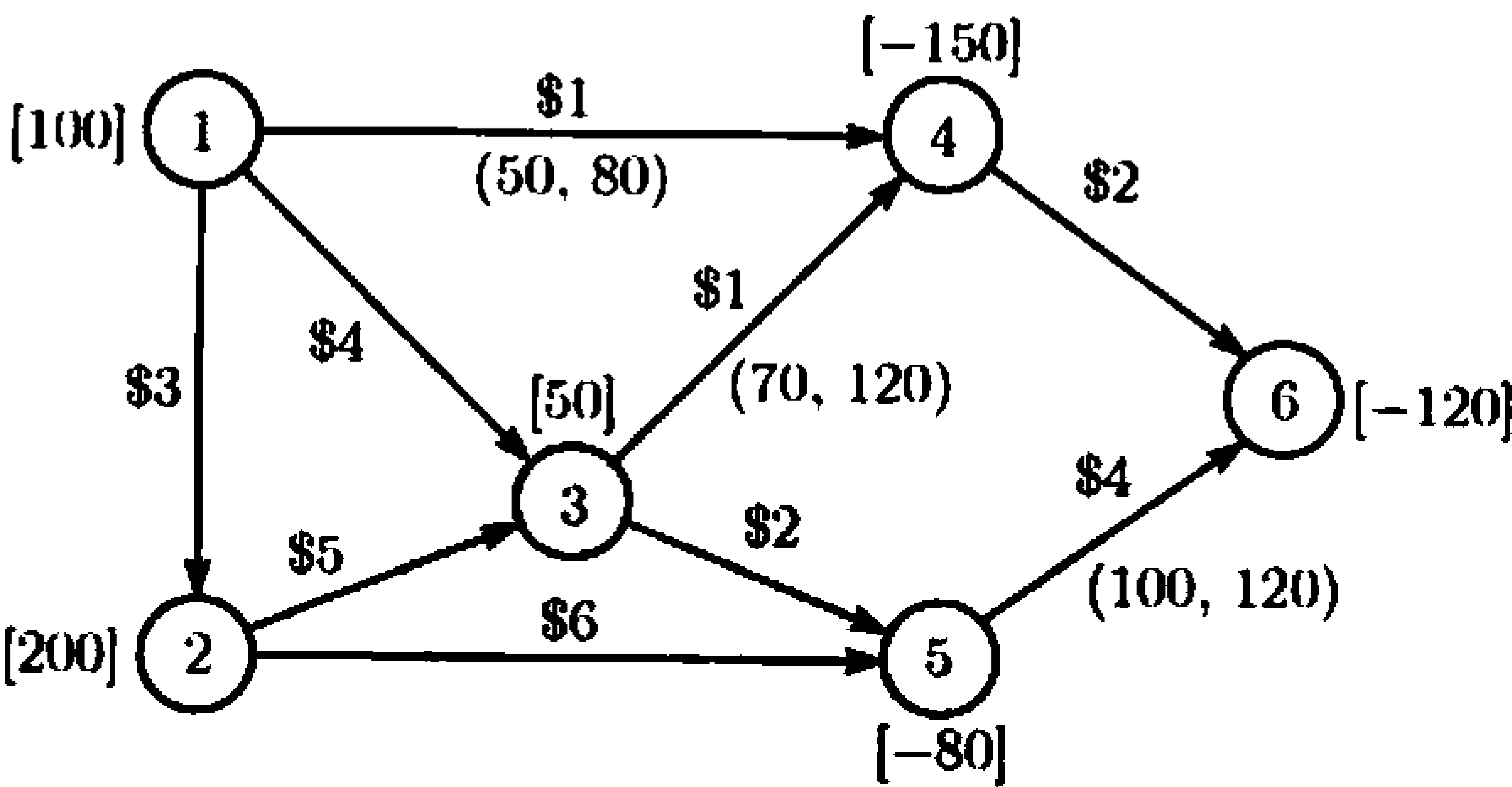


图 20.2　例 20.1-1 中的容量限制网络

习题 20.1A

*1. 根据下面的数据, 生产某种产品来满足一个 4 阶段计划期的需求:

阶段	需求量	单位产品成本 ($)	单位产品的存储费用 ($)
1	100	24	1
2	110	26	2
3	95	21	1
4	125	24	2

假定不允许延迟交货, 用网络模型表示这个问题.

2. 在第 1 题中, 假定允许延迟交货, 延迟交货一个阶段则一个单位产品惩罚 $1.5. 用网络模型表示这个问题.

3. 在第 1 题中, 假定 4 个阶段的生产能力分别为 110, 95, 125, 100, 也就是说如果不允许延迟交货则某些阶段的需求量就得不到满足. 假定延迟交货一个阶段则一个单位惩罚 $1.5. 用网络模型表示这个问题.

4. Daw Chemical 公司拥有两家工厂, 给两个客户生产一种基本的化学化合物, 两个客户的需求量每月分别是 660 吨和 800 吨. 工厂 1 每月的产量介于 400 到 800 吨, 工厂 2 的产量介于 450 到 900 吨. 工厂 1 和工厂 2 生产这种化合物的单位 (每吨) 成本分别是 $25 和 $28. 工厂的原材料由两个供应商提供, 承诺为工厂 1 每月提供至少 500 吨, 每吨 $200; 为工厂 2 每月提供至少 700 吨, 每吨 $210. 同时 Daw Chemical 公司还要负责承担原材料运进以及化合物运出的运输费用. 从供应商 1 运输 1 吨原材料到工厂 1 和工厂 2 的费用分别是 $10 和 $12; 同样供应商 2 运输 1 吨原材料到工厂 1 和工厂 2 的费用分别是 $9 和 $13. 从工厂 1 运输 1 吨的化合物到客户 1 和客户 2 的费用分别是 $3 和 $4; 从工厂 2 运输 1 吨的化合物到客户 1 和客户 2 的费用分别是 $5 和 $2. 假定 1 吨的原材料可以生产 1 吨的化合物, 请用网络模型表示这个问题.

5. 要求两所公立学校通过招收少数民族的学生来维持种族平衡. 在两所学校的招生人数中少数民族的学生要占 30% 到 40%. 非少数民族的学生住在两个社区, 少数民族的学生住在 3 个其他的社区, 从 5 个社区和两所学校之间的距离见下表:

学校	最大招生人数	从学校到社区的往返里程 (英里)				
		少数民族社区			非少数民族社区	
		1	2	3	1	2
1	1 500	20	12	10	4	5
2	2 000	15	18	8	6	5
	学生总数	500	450	300	1 000	1 000

建立这个问题的网络模型, 并确定每个学校的少数民族学生和非少数民族学生招生的人数.

6. (Charnes and Cooper, 1967) 在接下来的 3 个月里, 农场主们收割庄稼并卖给一个当地的收购中心, 然后收购中心再将粮食出售给零售商. 下表给出了收购中心在这 3 个月中收购粮食和出售粮食的价格.

月份	每吨的收购价格 ($)	每吨的出售价格 ($)
1	200	250
2	190	230
3	195	240

收购中心的最大容量是 800 吨, 可以用于购买新粮食的现金有 $100 000. 第 1 个月开始的时候, 收购中心的容量已经占用了一半. 估计收购中心每月每吨粮食的存储费用为 $10. 剩余的现金每月有 1% 的利息. 目标是确定一个购买/出售策略, 使得在第 3 个月结束的时候可以获得最多的现金.

建立这个问题的网络模型.

20.1.2　线性规划模型

我们为带有容量限制的网络模型建立线性规划模型, 它是 20.1.3 节将要涉及的限制单纯形算法的基础. 使用 20.1.1 节引入的符号, 可以为容量限制网络建立下面的线性规划:

$$\min\ z = \sum_{(i,j)\in A}\sum c_{ij}x_{ij}$$

$$\text{s.t.}\ \sum_{\substack{k\\(j,k)\in A}} x_{jk} - \sum_{\substack{i\\(i,j)\in A}} x_{ij} = f_j,\quad j\in N$$

$$l_{ij} \leqslant x_{ij} \leqslant u_{ij}, (i,j)\in A$$

满足节点 j 的网络流 f_j 的等式为:

$$\text{节点 } j \text{ 的输出流} - \text{节点 } j \text{ 的输入流} = f_j$$

如果 $f_j > 0$, 那么节点 j 被视为一个源点; 如果 $f_j < 0$ 则被视为一个汇点.

利用下面的式子进行代换, 可以把变量的下界去掉:

$$x_{ij} = x'_{ij} + l_{ij}$$

新的网络流变量 x'_{ij} 有上界约束 $u_{ij} - l_{ij}$, 同时节点 i 的网络流变为 $f_i - l_{ij}$, 节点 j 的网络流变为 $f_j + l_{ij}$. 图 20.3 给出了对于弧 (i,j) 进行下界替换的过程.

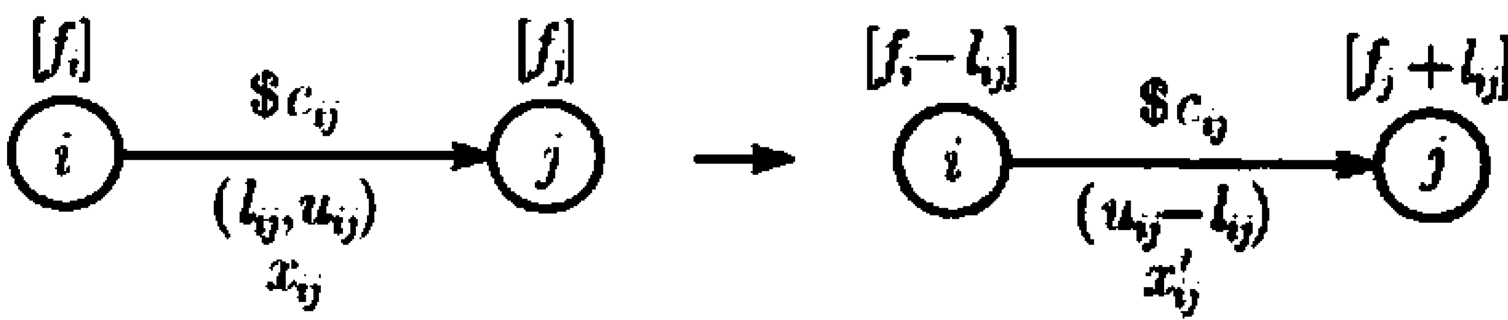


图 20.3　去掉弧上的下界约束

例 20.1-2

写出图 20.2 所表示的网络的弧上下界约束代换之前和之后的线性规划模型.

根据每个节点上的输入和输出流的关系得到线性规划的主要约束，下面是相应的 LP 模型：

min	x_{12}	x_{13}	x_{14}	x_{23}	x_{25}	x_{34}	x_{35}	x_{46}	x_{56}	
	3	4	1	5	6	1	2	2	4	
节点 1	1	1	1							= 100
节点 2	−1			1	1					= 200
节点 3		−1		−1		1	1			= 50
节点 4			−1			−1		1		=−150
节点 5					−1		−1		1	= −80
节点 6								−1	−1	=−120
下界	0	0	50	0	0	70	0	0	100	
上界	∞	∞	80	∞	∞	120	∞	∞	120	

留心观察约束系数可以发现，对应变量 x_{ij} 的列中恰好在第 i 行有一个 1，在第 j 行有一个 −1，这也正是网络模型的一个典型特点. 剩余的变量的系数都是 0. 这种结构是网络流模型的一个典型特征.

最优值是 $z = \$1\,870\,000$，其中相应的最优解为 $x_{13} = 20$(千蒲式耳)，$x_{14} = 80$，$x_{23} = 20$，$x_{25} = 180$，$x_{34} = 90$，$x_{46} = 20$，$x_{56} = 100$，其余变量等于 0.

带有下界约束的变量可以使用下面的代换：

$$x_{14} = x'_{14} + 50$$

$$x_{34} = x'_{34} + 70$$

$$x_{56} = x'_{56} + 100$$

利用上面的代换，可以得到下面的线性规划模型：

min	x_{12}	x_{13}	x'_{14}	x_{23}	x_{25}	x'_{34}	x_{35}	x_{46}	x'_{56}	
	3	4	1	5	6	1	2	2	4	
节点 1	1	1	1							= 50
节点 2	−1			1	1					= 200
节点 3		−1		−1		1	1			= −20
节点 4			−1			−1		1		= −30
节点 5					−1		−1		1	=−180
节点 6								−1	−1	= −20
上界	∞	∞	30	∞	∞	50	∞	∞	20	

经过下界代换之后的网络如图 20.4 所示. 值得注意的是，可以通过使用图 20.3 中弧上的下界代换直接得到图 20.4 的网络，而不需要首先将问题表示为一个线性规划模型.

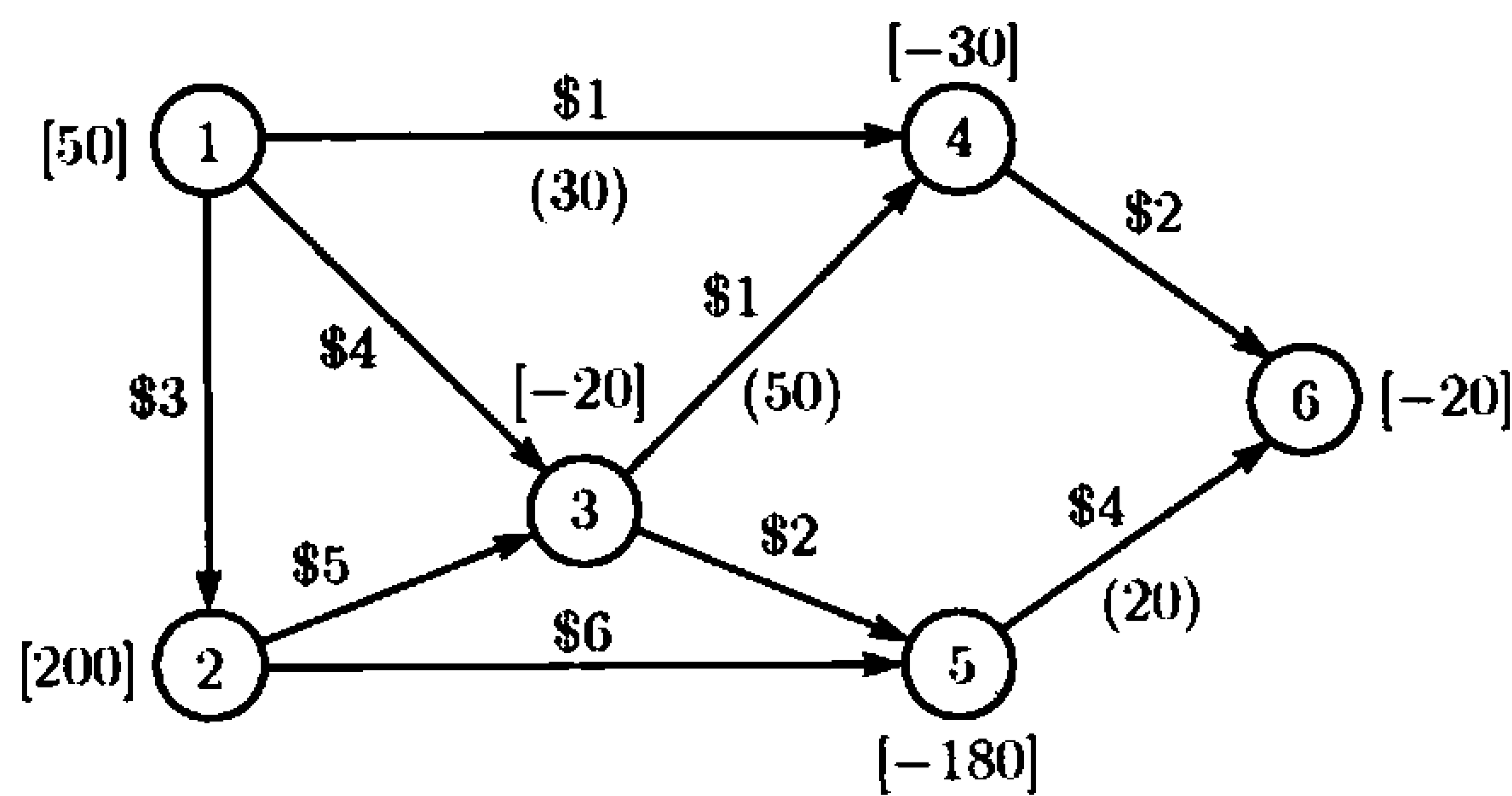


图 20.4 例 20.1-2 中经过下界代换之后的网络

最优值是 $z' = \$1\ 350\ 000$(或者 $z = 1\ 350\ 000 + 50 \times 1\ 000 + 70 \times 1\ 000 + 100 \times 4\ 000 = \$1\ 870\ 000$), 其中最优解是 $x_{13} = 20$(千蒲式耳), $x'_{14} = 30$(或者 $x_{14} = 30 + 50 = 80$), $x_{23} = 20$, $x_{25} = 180$, $x'_{34} = 20$(或者 $x_{34} = 20 + 70 = 90$), $x_{46} = 20$, $x'_{56} = 0$(或者 $x_{56} = 0 + 100 = 100$). 这与进行下界代换之前求出的最优解相同.

例 20.1-3 (雇用调度)

下面给出的例题的网络模型将要说明, 虽然初始的时候不满足节点上的流的需求 (即不满足节点的输出流减去输入流等于这个节点网络流), 但是, 可以通过对线性规划模型的约束进行一些相应操作, 使得节点上的流满足网络流模型的要求.

Tempo Employment Agency 拿到一份合同, 需要在接下来的 4 个月 (1 月到 4 月) 按照下面的要求组织工人生产:

月份	1 月	2 月	3 月	4 月
工人数目	100	120	80	170

根据每月工人人数需求的变化, 或许让工人连续工作会比分阶段雇用工人更节约. 下表给出了雇用一名工人的费用取决于这名工人的工作月数:

雇用时间长度 (月数)	1	2	3	4
每名工人的费用 ($)	100	130	180	220

令

$$x_{ij} = \text{从 } i \text{ 月份初雇用到 } j \text{ 月份初结束的工人数}$$

例如, x_{12} 表示雇用只在 1 月份工作的工人数. 为了建立这个问题的线性规划模型, 我们引入 5 月作为虚拟月份, 则 x_{45} 表示雇用只在 4 月份工作的工人数.

根据变量的定义, 对于所有满足 $i \leqslant k < j$ 的 x_{ij}, 都可以来满足 k 月份的工人数需要. 用 $s_i \geqslant 0$ 表示 i 月份的剩余工人数, 那么线性规划模型是:

min	x_{12}	x_{13}	x_{14}	x_{15}	x_{23}	x_{24}	x_{25}	x_{34}	x_{35}	x_{45}	s_1	s_2	s_3	s_4	
	100	130	180	220	100	130	180	100	130	100					
1 月	1	1	1	1							−1				=100
2 月		1	1	1	1	1	1					−1			=120
3 月			1	1		1	1	1	1				−1		= 80
4 月				1			1		1	1				−1	=170

上面得到的线性规划模型不是一个具有 $(-1,1)$ 特殊结构的网络流模型 (参见例 20.1-2), 但是, 如果使用下面的算术操作就可以将上面的线性规划转化为一个网络流模型.

(1) 对于带有 n 个等式的线性规划, 引入一个新的等式, 记作第 $n+1$ 个, 它是将第 n 个等式两边同时乘以 -1 得到的.

(2) 第 1 个等式不变.

(3) 对于 $i=2,3,\cdots,n$, 用 (等式 i)−(等式 $i-1$) 来代替等式 i.

将上面的操作应用到本例题的雇用调度中, 可以得到下面的线性规划模型, 这个模型的结构满足网络流模型:

min	x_{12}	x_{13}	x_{14}	x_{15}	x_{23}	x_{24}	x_{25}	x_{34}	x_{35}	x_{45}	s_1	s_2	s_3	s_4	
	100	130	180	220	100	130	180	100	130	100					
1 月	1	1	1	1							−1				= 100
2 月	−1				1	1	1				1	−1			= 20
3 月		−1			−1			1	1			1	−1		= −40
4 月			−1			−1		−1		1			1	−1	= 90
5 月				−1			−1		−1	−1				1	=−170

使用上面的线性规划模型, 可以将雇用调度问题等价地表示成如图 20.5 所示的最小费用流的网络. 由于这个问题中弧上没有容量上界的约束, 所以它也可以看作一个运输模型 (参见 5.5 节).

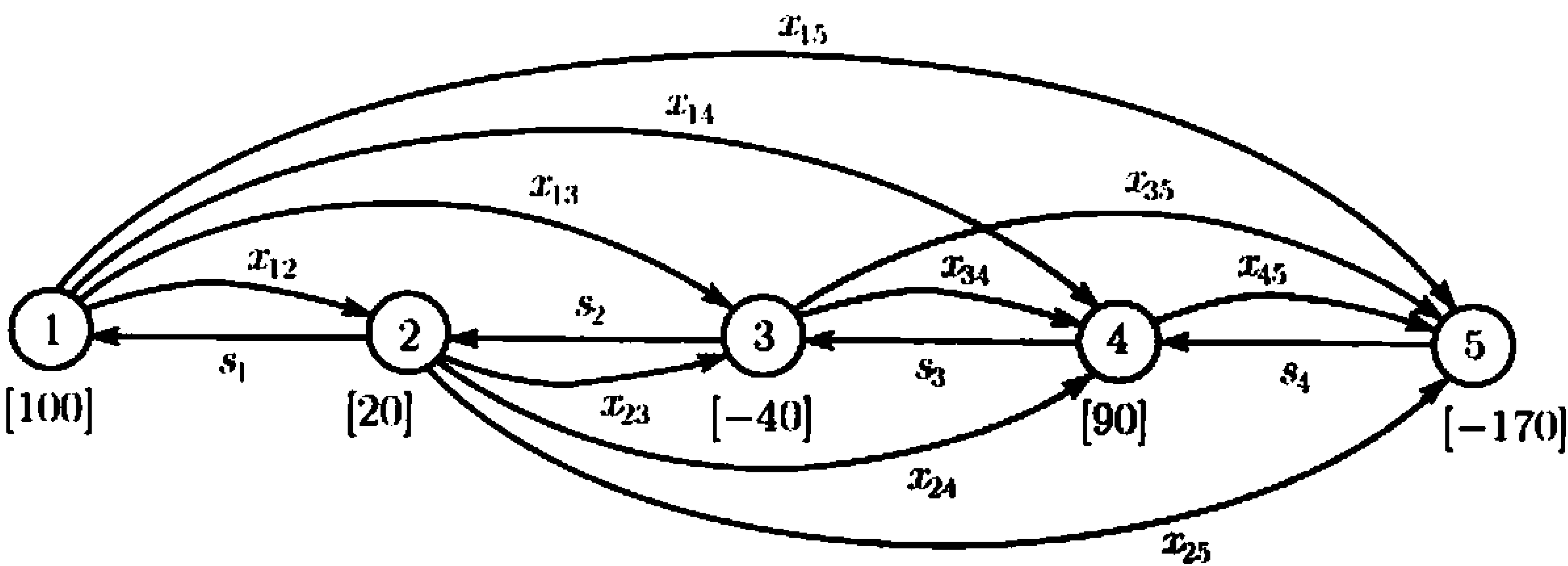


图 20.5 雇用调度问题的网络模型

最优解是 $x_{15}=100$, $x_{25}=20$, $x_{45}=50$, 其他的变量都是 0. 下表给出了这 4 个月份中雇用和解雇的情况. 总的费用是 \$30 600.

月份	雇用的工人数	解雇的工人数	工人总数
1 月	100	0	100
2 月	20	0	120
3 月	0	0	120
4 月	50	0	170

习题 20.1B

*1. 写出图 20.6 所表示的最小费用流网络在经过下界代换之前和之后的线性规划模型.

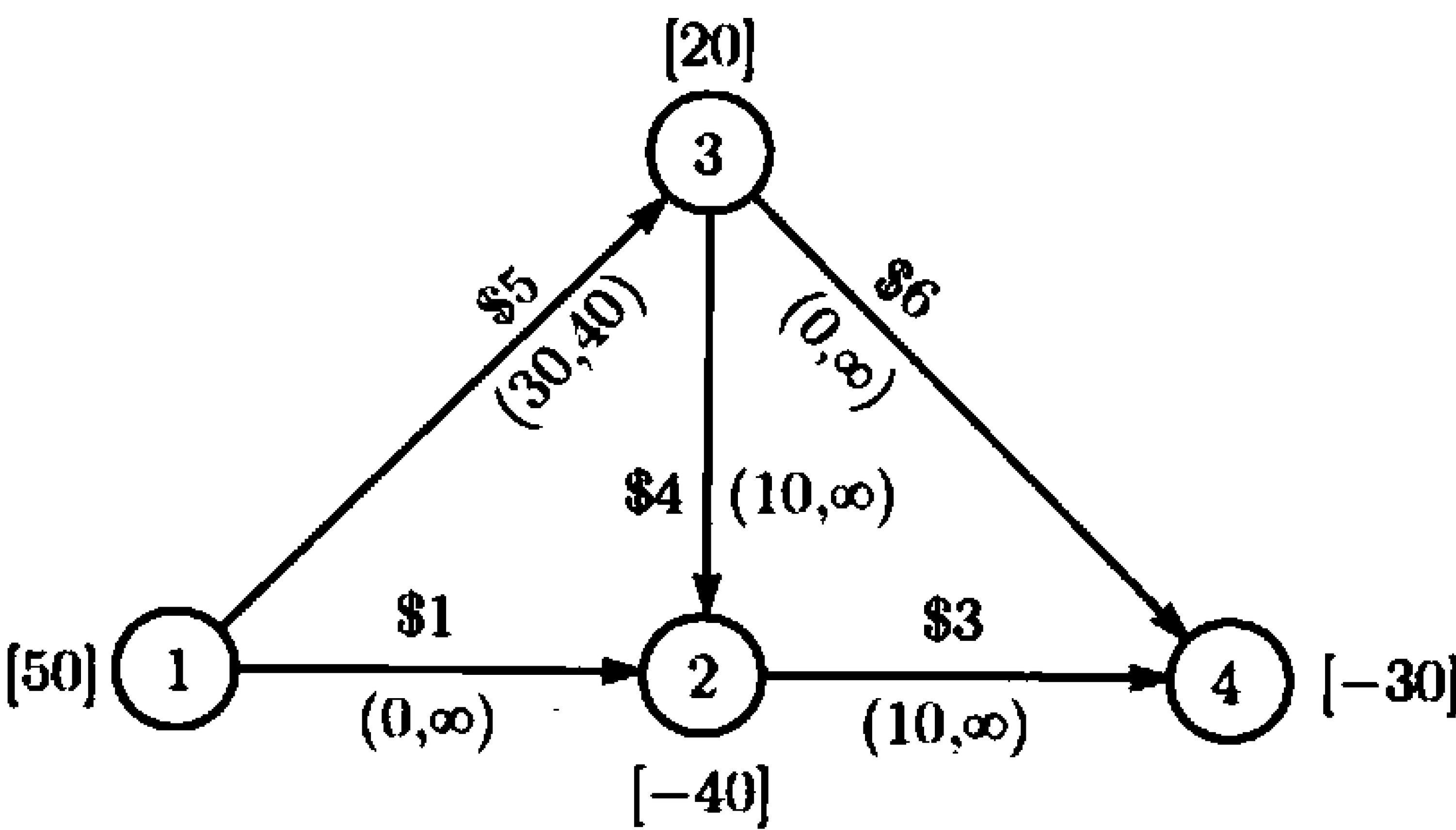


图 20.6　习题 20.1B 第 1 题的网络

2. 使用检查法, 找到例 20.1-3(图 20.5) 中雇用调度问题最小费用流模型的一个可行解. 说明这个解给出的雇用和解雇的模式满足每一个月的工人数需求, 并计算出相应的总费用.
3. 在例 20.1-3 中, 假定一个工人至少要被雇用 2 个月, 重新建立雇用调度模型. 写出相应的线性规划, 并把它转化成一个最小费用网络流模型.
4. 对于例 20.1-3 中雇用调度的模型, 根据下面给定的 5 个月工人数需求数据, 建立线性规划模型和相应的最小费用流的网络. 雇用一名工人连续工作 1 到 5 个月的每月费用为 \$50, \$70, \$85, \$100, \$130.

   (a)

月　　份	1	2	3	4	5
需要的工人数	300	180	90	170	200

   (b)

月　　份	1	2	3	4	5
需要的工人数	200	220	300	50	240

5. 将有容量限制的网络转化为无容量限制的网络.证明: 一条带有容量限制 $x_{ij} \leqslant u_{ij}$ 的弧 $(i \to j)$ 可以用两条无容量限制的弧 $(i \to k)$ 和 $(k \to j)$ 来代替, 其中在节点 k 上增加一个输出, 流 $[-u_{ij}]$, 在节点 j 上增加一个输入流 $[+u_{ij}]$. 这样做的结果就是将一个有容量限制的网络转化成一个无容量限制的运输模型 (参见 5.1 节). 将上面的结论应用到图 20.7 给出的网络上, 并通过求解无容量限制的运输模型来找出原始网络的最优解.

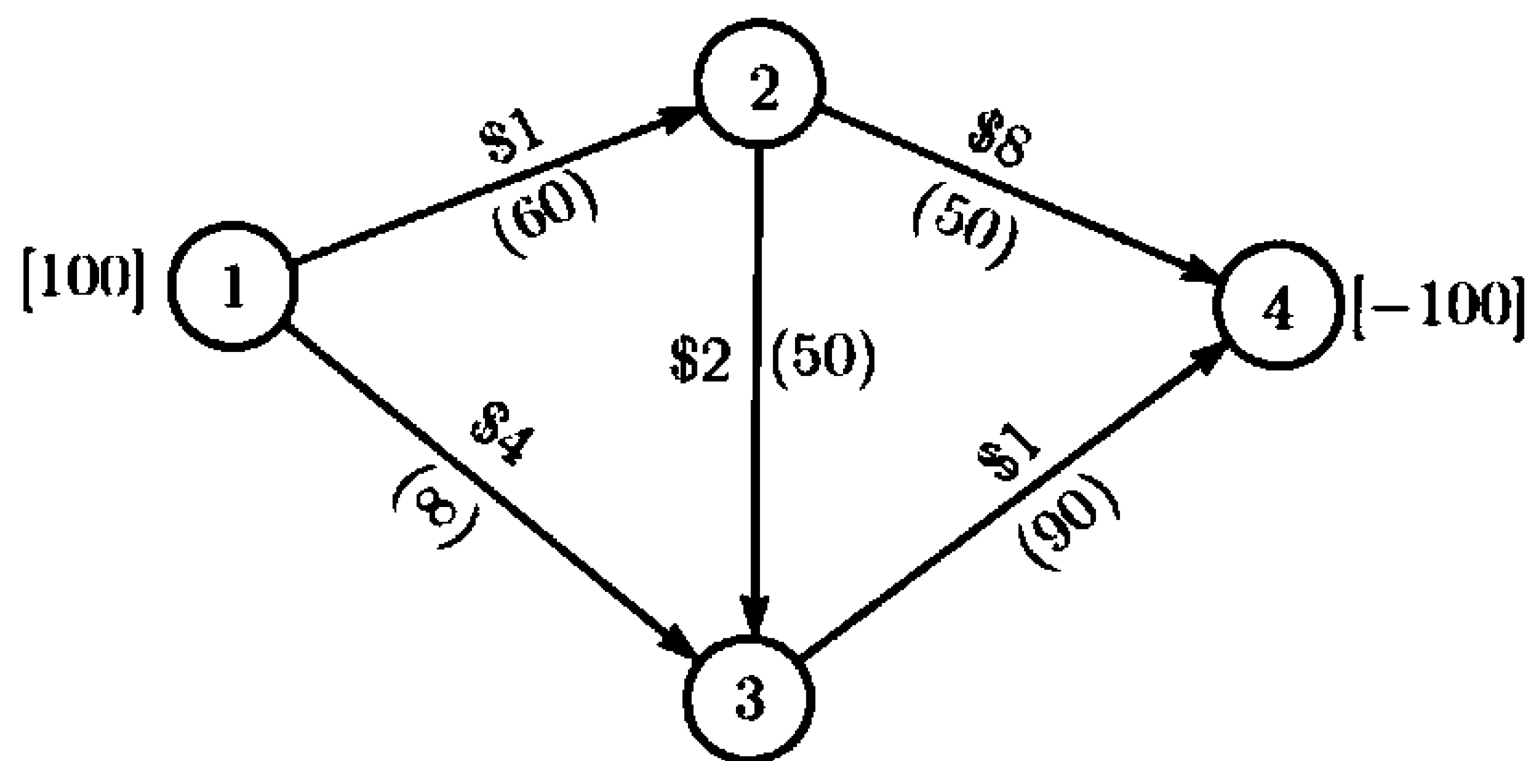


图 20.7 习题 20.1B 第 5 题的网络

20.1.3 带有容量限制的网络的单纯形算法

这种算法的步骤仍然是经典的单纯形方法, 只不过是为了适应最小费用流模型的特殊网络结构而设计的.

根据 20.1.2 节定义的线性规划模型, 给定节点 i 上的网络流为 f_i, 有容量限制的单纯形算法要求网络满足

$$\sum_{i=1}^{n} f_i = 0$$

这个条件告诉我们, 网络中的总供应量等于总需求量. 当这两个量不相等时, 我们总可以满足这个条件, 通过增加一个平衡虚拟源点或者汇点, 将这个点与网络中所有的点都用弧相连, 并且弧上没有容量限制, 单位流费用是 0. 然而, 网络流量平衡不一定就能保证网络存在可行解, 因为是否有可行解依赖于弧上的容量限制.

下面给出有容量限制的单纯形算法的步骤. 当然首先要熟悉单纯形方法和对偶理论 (参见第 3 章和第 4 章), 同时还要了解上有界的单纯形方法 (参见 13.3 节).

第 0 步 找到网络的一个初始的基本可行解 (弧的集合). 进入第 1 步.

第 1 步 使用单纯形方法最优性条件, 确定一个进入基的弧 (变量). 如果解是最优的, 停止; 否则, 进入第 2 步.

第 2 步 使用单纯形方法可行性条件, 确定一个离开基的弧 (变量). 得到一个新的解, 然后进入第 1 步.

对于一个网络流为 0(即 $f_1 + f_2 + \cdots + f_n = 0$) 的 n−节点网络, 有 $n-1$ 个*独立* 的约束等式. 因此, 一个基本可行解一定包含 $n-1$ 条弧. 可以证明一个基本可行解恰好对应网络上的一棵*生成树*(参见 6.2 节).

第 1 步中的进基弧可根据计算当前所有非基弧 (i, j) 的目标系数来确定. 如果所有的目标系数都 $\leqslant 0$, 那么当前的基本可行解就是最优的. 否则, 选择目标系数为最大正数的非基弧进入基.

计算目标系数是基于对偶性, 恰好与我们在运输模型中所作的一样 (见 5.3.2 节). 利用 20.1.2 节定义的线性规划模型, 令 w_i 表示与节点 i 的约束相对应的对偶

变量, 那么对偶问题 (排除上界约束) 是:

$$\max \ z=\sum_{i=1}^{n} f_i w_i$$
$$\text{s.t.} \ w_i - w_j \leqslant c_{ij}, \quad (i,j)\in A$$
$$w_i \text{ 没有符号限制}, \quad i=1,2,\cdots,n$$

根据线性规划理论, 我们得到

$$w_i - w_j = c_{ij}, \text{ 对于基弧 } (i,j)$$

因为原始的线性规划 (20.1.2 节) 根据定义有一个多余的约束, 我们可以给一个对偶变量分配一个任意值 (与 5.3 节中的运输算法进行比较). 为简便起见, 令 $w_1=0$. 然后可以根据基等式 $w_i - w_j = c_{ij}$ 来确定剩余的对偶变量的值. 根据 4.2.4 节的公式 2 可以知道, 非基变量 x_{ij} 的目标系数是相应的对偶约束左右两边的差值, 即 $w_i - w_j - c_{ij}$. 令

$$z_{ij} = w_i - w_j$$

变量 x_{ij} 的目标系数是

$$z_{ij} - c_{ij} = w_i - w_j - c_{ij}$$

还需要确定剩余的变量, 下面的例题给出了详细的过程.

例 20.1-4

现在有一个连接两个海水淡化厂与两个城市的管道网络, 两个工厂每天的供应量分别为 40 和 50 百万加仑, 城市 1 和城市 2 每天的需求量分别是 30 和 60 百万加仑. 节点 1 和节点 2 分别代表工厂 1 和工厂 2, 节点 4 和节点 5 分别代表城市 1 和城市 2, 节点 3 表示工厂与城市之间的一个升压站. 因为节点 1 和节点 2 的总供应量等于节点 4 和节点 5 的总需求量, 所以这个模型已经是平衡的. 图 20.8 给出了相应的网络.

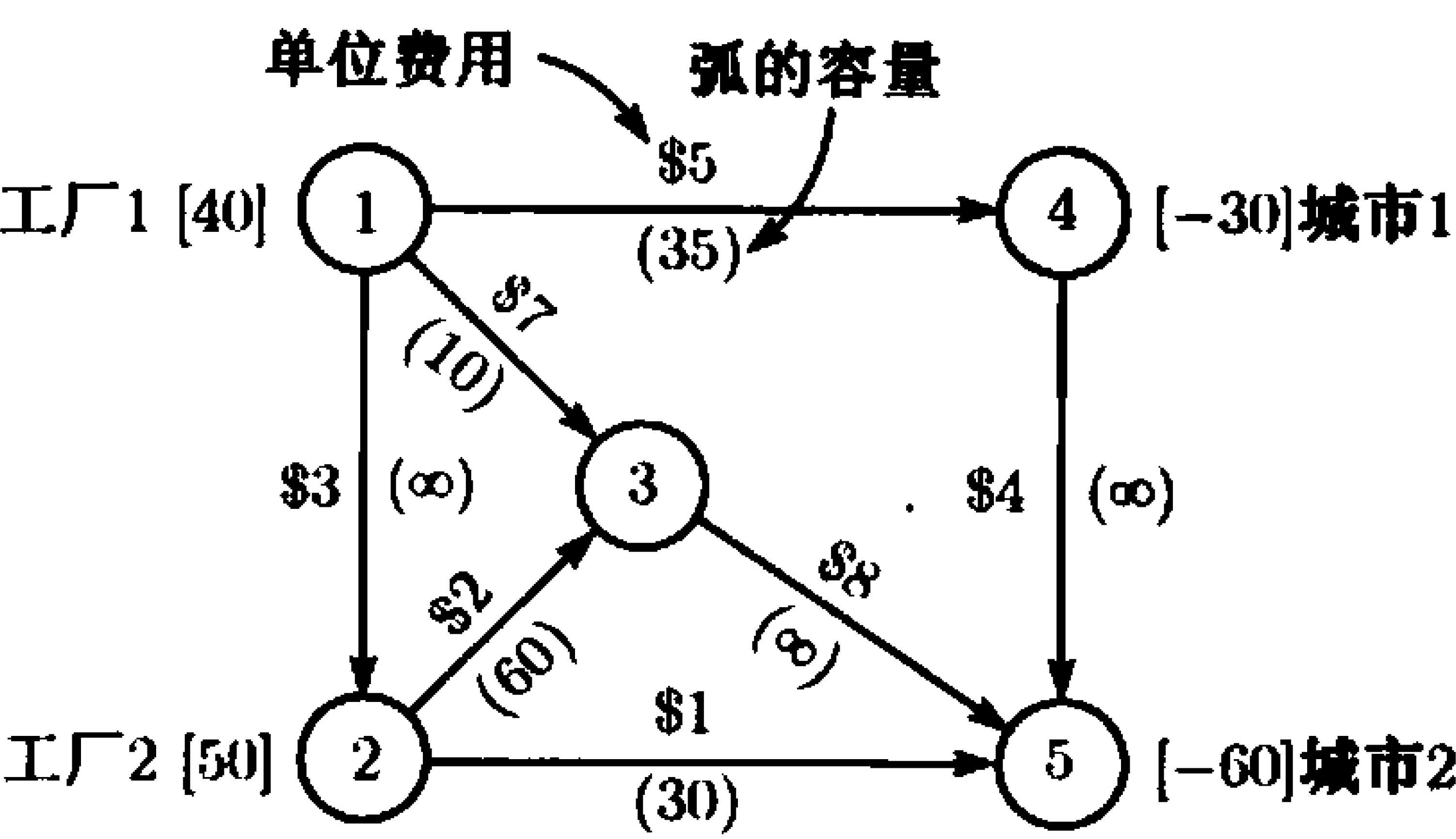


图 20.8 例 20.1-4 的网络

迭代 0

第 0 步 *确定一个初始的基本可行解.* 初始的基本可行解一定是一棵生成树. 通过直观的检查可以得到一棵可行的生成树, 如图 20.9(图中粗黑的弧) 所示. 一般的我们使用人工的方法得到这样一个可行解 (参见 Bazaraa 等人, 1990, pp. 440-446).

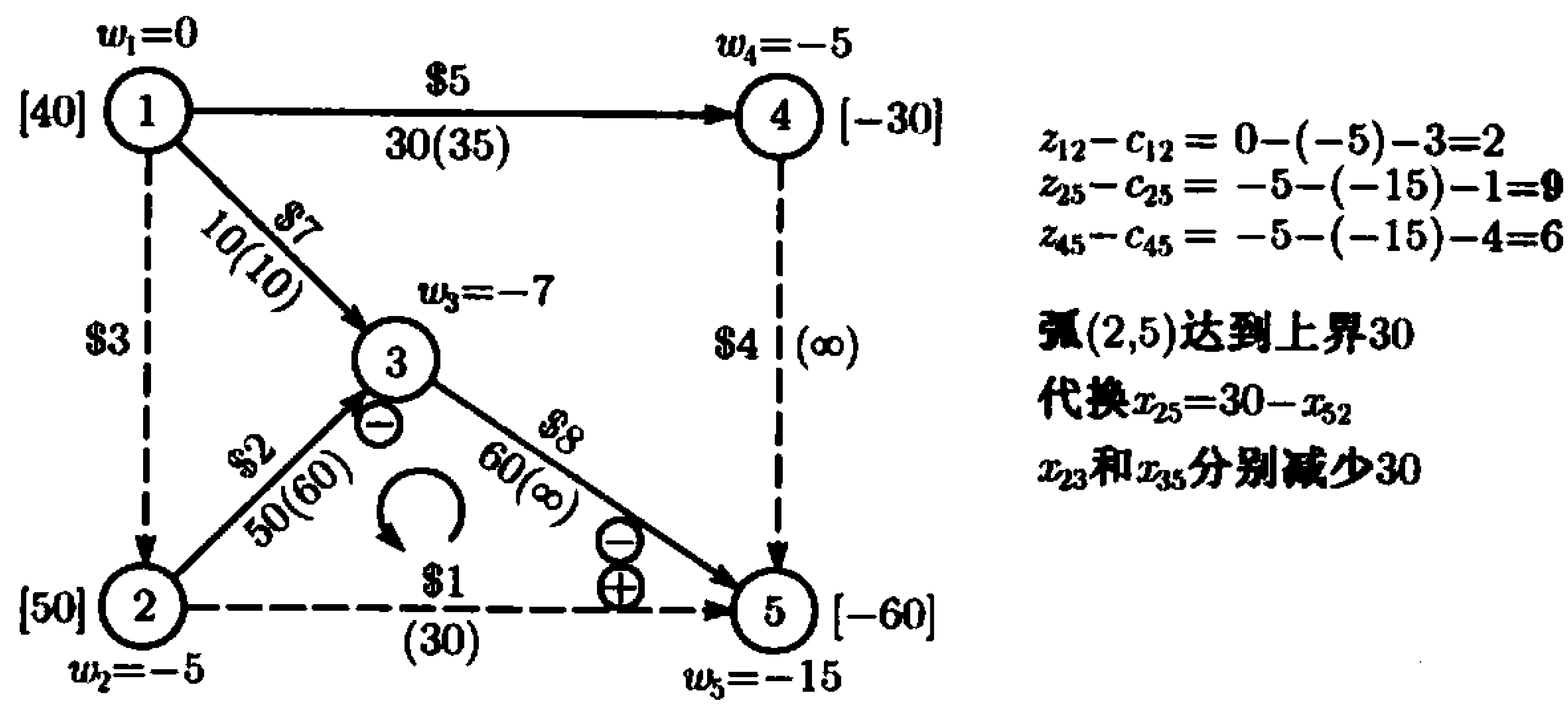


图 20.9 第 0 次迭代后的网络

在图 20.9 中, 基本可行解是由粗弧 $(1,3),(1,4),(2,3),(3,5)$ 组成的, 可行的流量分别为 $30,10,50,60$. 剩余的弧 $(1,2),(2,5),(4,5)$ 都表示非基变量. 弧上的符号 $x(c)$ 的标记表示相应的弧上的流量为 x 个单位, 这条弧上的容量为 c. x 和 c 的预设值分别是 0 和 ∞.

迭代 1

第 1 步 *确定进基的弧.* 通过求解当前的基方程可以得到对偶变量的值:

$$w_1 = 0$$

$$w_i - w_j = c_{ij}, \quad \text{对于所有的基弧}(i,j)$$

于是得到

$$\begin{aligned}
&\text{弧 (1,3):} && w_1 - w_3 = 7, && \text{因此} w_3 = -7\\
&\text{弧 (1,4):} && w_1 - w_4 = 5, && \text{因此} w_4 = -5\\
&\text{弧 (2,3):} && w_2 - w_3 = 2, && \text{因此} w_2 = -5\\
&\text{弧 (3,5):} && w_3 - w_5 = 8, && \text{因此} w_5 = -15
\end{aligned}$$

下面为非基弧计算 $z_{ij} - c_{ij}$.

$$\begin{aligned}
&\text{弧 (1,2):} && w_1 - w_2 - c_{12} = 0 - (-5) - 3 = 2\\
&\text{弧 (2,5):} && w_2 - w_5 - c_{25} = (-5) - (-15) - 1 = 9\\
&\text{弧 (4,5):} && w_4 - w_5 - c_{45} = (-5) - (-15) - 4 = 6
\end{aligned}$$

因此, 取弧 (2,5) 进入基本可行解.

第 2 步 ***确定离开基的弧***. 从图 20.9 可以看出, 弧 (2,5) 与基弧 (2,3) 和 (3,5) 构成一个环. *根据生成树的定义可以知道, 不允许再有其他的环*. 由于我们需要增加新弧 (2,5) 上的流量, 所以必须调整这个环上其他弧上的流量来保证新的解是可行的. 要做到这一点, 我们给这个环上增加的弧 (就是从节点 2 到节点 5 的弧) 标记一个带有 (+) 的流来代表这个流的方向. 然后对环上剩余的弧, 根据它与新增加弧的流的方向*相同或不同*, 也标记上 (+) 或者 (−). 图 20.9 给出了这种标记的过程.

根据下面的两个条件来确定进基弧 (2,5) 上可以增加的最大流量:

(1) 环上包含的当前基弧上的新流量不能为负数.

(2) 进基弧上新的流量不能超过它的容量.

应用第 (1) 个条件可以知道, 弧 (2,3) 和 (3,5) 上的流量减少不能超过 $\min\{50, 60\} = 50$ 个单位. 根据第 (2) 个条件, 弧 (2,5) 上的流量增加不能超过它的容量 (= 30 个单位). 因此在这个环上的流量最多可以改变 $\min\{30, 50\} = 30$ 个单位, 修改之后的流量分别是: 弧 (2,5) 上 30 个单位, 弧 (2,3) 上 $50 - 30 = 20$ 个单位, 弧 (3,5) 上 $60 - 30 = 30$ 个单位.

由于当前的基本可行解的弧中没有流量为 0 的弧离开基, 因此新的弧 (2,5) 必须仍然为非基弧, 并且流量达到了它的上界. 然而为了避免处理这种已经达到容量上界的非基弧, 我们引入下面的代换:

$$x_{25} = 30 - x_{52},\ 0 \leqslant x_{52} \leqslant 30$$

引入这个代换就会影响节点 2 和节点 5 上的流量方程. 考虑下面的方程.

节点 2 上当前的流量方程: $50 + x_{12} = x_{23} + x_{25}$

节点 5 上当前的流量方程: $x_{25} + x_{35} + x_{45} = 60$

因此, 利用代换 $x_{25} = 30 - x_{52}$, 可以得到

节点 2 上新的流量方程: $20 + x_{12} + x_{52} = x_{23}$

节点 5 上新的流量方程: $x_{35} + x_{45} = x_{52} + 30$

经过代换后的结果如图 20.10 所示. 现将弧 (2,5) 上流的方向改为 $5 \to 2$, 同时在代换过程中要将弧 (2,5) 上单位流的费用改为 −$1. 通过在相应弧上的容量旁边标记一个星号来表明已对这条弧进行了反向.

迭代 2

图 20.10 给出了新的 $z_{ij} - c_{ij}$ 值 (请验证!), 并且指明了弧 (4,5) 进入基本可行解. 它还定义了与新弧的进入有关的环, 并对环上的弧进行了符号标记.

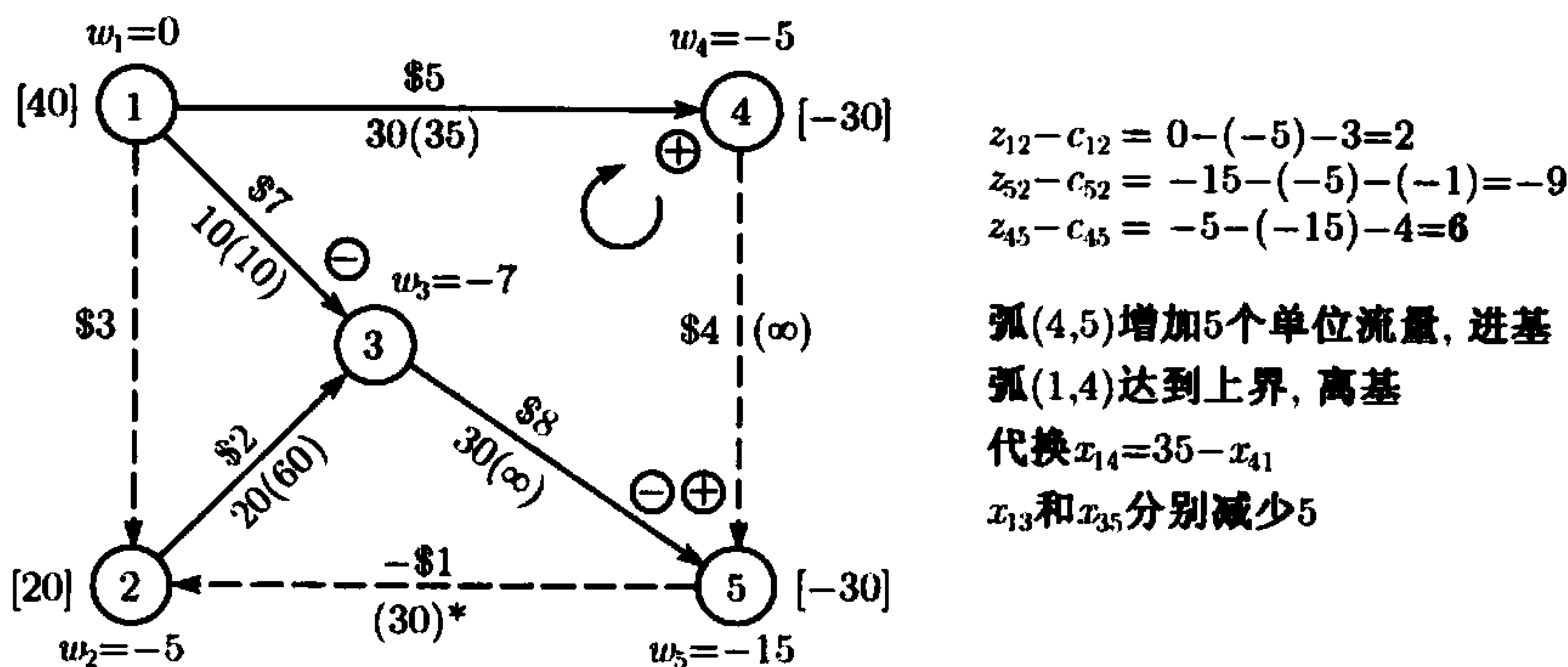


图 20.10 第 1 次迭代后的网络

根据下面几个量的最小值来增加弧 (4, 5) 的流量.

(1) 弧 (4,5) 上最大可以允许*增加*的流量是 ∞.

(2) 弧 (1,4) 上最大可以允许*增加*的流量是 35 − 30 = 5 个单位.

(3) 弧 (1,3) 上最大可以允许*减少*的流量是 10 个单位.

(4) 弧 (3,5) 上最大可以允许*减少*的流量是 30 个单位.

因此, 弧 (4,5) 上最多可以增加 5 个单位的流量, 这使得弧 (4,5) 成为进基弧, 同时弧 (1,4) 在上界 (= 35) 处成为非基弧.

引入代换 $x_{14} = 35 - x_{41}$, 那么网络将变成如图 20.11 所示, 弧 (1,3), (2,3), (3,5), (4,5) 组成基本可行解 (生成树). 同时弧 (1,4) 上的反向流的单位费用变为 −\$5. 节点 1 和节点 4 的流量方程也要随之代换, 每个节点有 5 个单位的输入.

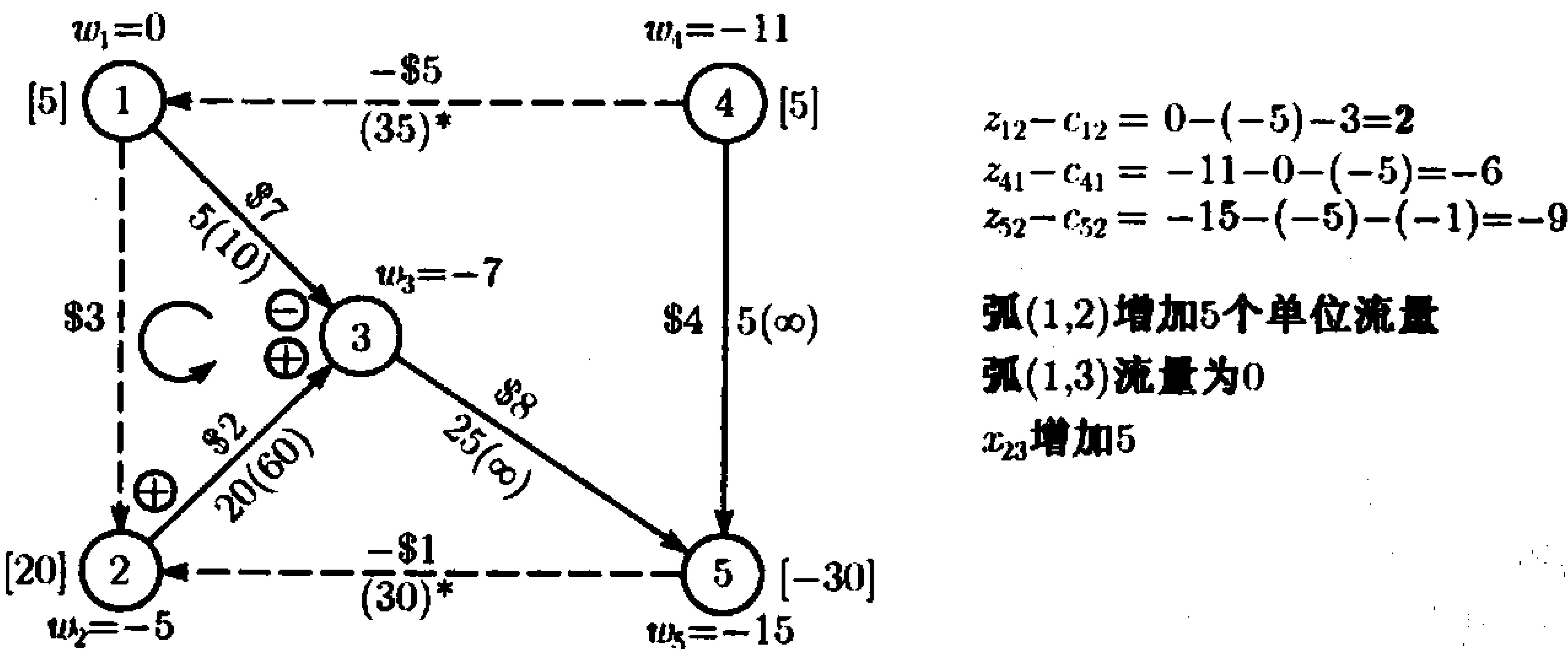


图 2.11 第 2 次迭代后的网络

迭代 3

图 20.11 给出了非基弧 (1,2), (4,1), (5,2) 的新的 $z_{ij} - c_{ij}$ 值, 并且选择弧 (1,2) 进基, 流量增加 5 个单位, 同时弧 (1,2) 成为非基弧, 流量变为 0. 图 20.12 给出了新的解.

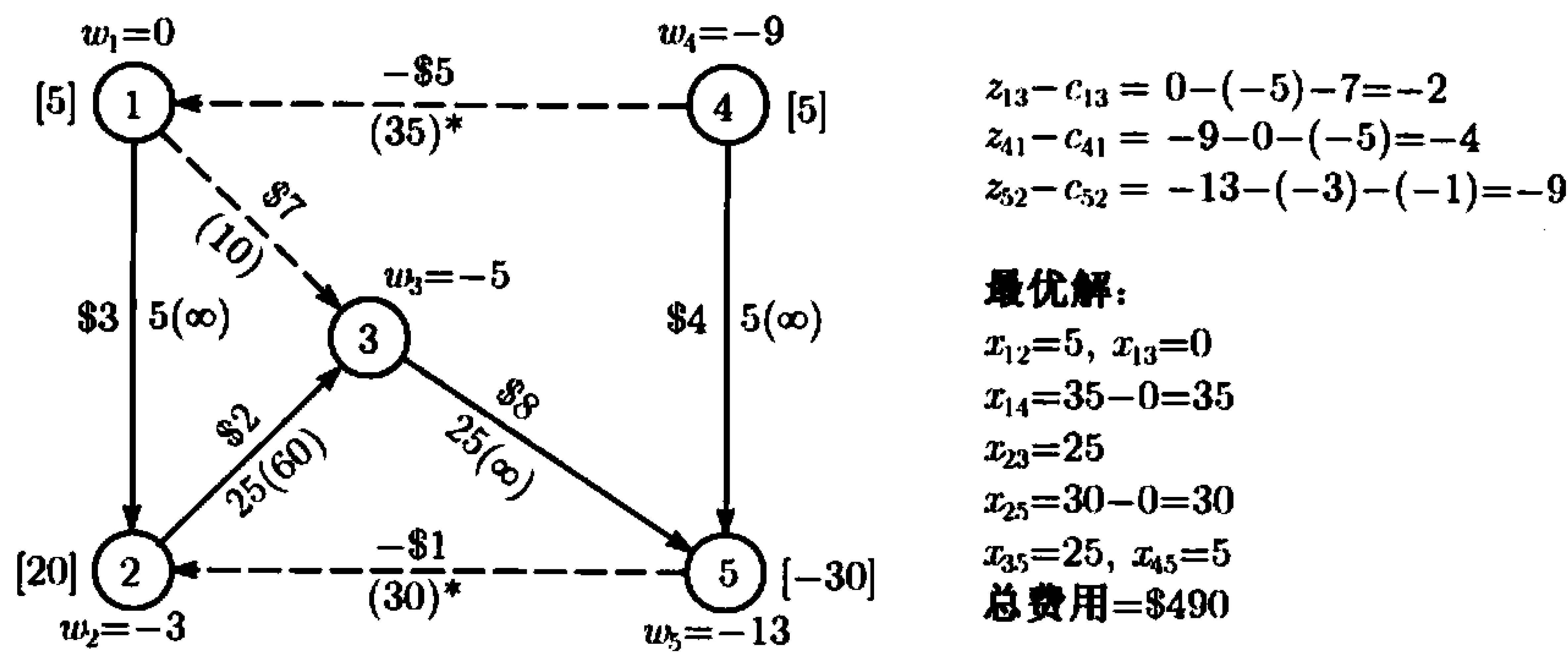


图 20.12　第 3 次迭代后的网络

迭代 4

图 20.12 中标出的新的 $z_{ij}-c_{ij}$ 值表明当前的解是最优解. 将变量反向代换就可以得出原始变量值, 如图 20.12 所示.

Excel 规划求解程序

Excel 规划求解关于解带有容量限制的最小费用流模型的基本思想, 与例题 6.3-6 求解最短路径模型的方法类似. 文件 solverEx20.1-4.xls 给出了详细的描述.

AMPL 程序

文件 amplEx20.1-4.txt 给出了求解任何规模的带有容量限制的最小费用流问题的一般模型. 模型的思想与例题 6.3-6 求解最短路径 AMPL 模型的方法类似.

习题 20.1C

*1. 利用容量限制的单纯形算法求解习题 20.1A 第 1 题, 并证明它也可以用运输模型来求解.
2. 利用容量限制的单纯形算法求解习题 20.1A 第 2 题, 并证明它也可以用运输模型来求解.
3. 利用容量限制的单纯形算法求解习题 20.1A 第 3 题.
4. 利用容量限制的单纯形算法求解习题 20.1A 第 4 题.
*5. 利用容量限制的单纯形算法求解习题 20.1A 第 5 题.
6. 利用容量限制的单纯形算法求解例 20.1-3 的雇用调度问题.
7. 怀俄明电力公司通过使用现有的泥浆管道来运输煤炭 (通过向管道中注水完成运输), 将煤炭从 3 个开采区 (1, 2, 3) 运输到 3 个发电厂 (4, 5, 6). 每一段管道每小时最多可以运输 10 吨, 下表给出了每吨的运输费用, 以及每小时的供应和需求.

	4	5	6	供应量
1	\$5	\$8	\$4	8
2	\$6	\$9	\$12	10
3	\$3	\$1	\$5	18
需求量	16	6	14	

求出最优的运输调度.

8. 图 20.13 中的网络给出了 7 个城市之间的距离. 利用容量限制的单纯形算法求出城市 1 到城市 7 的最短路径的距离. (提示: 可以假定城市 1 和城市 7 的网络流量分别为 [+1] 和 [−1], 其他节点的网络流量为 0.)

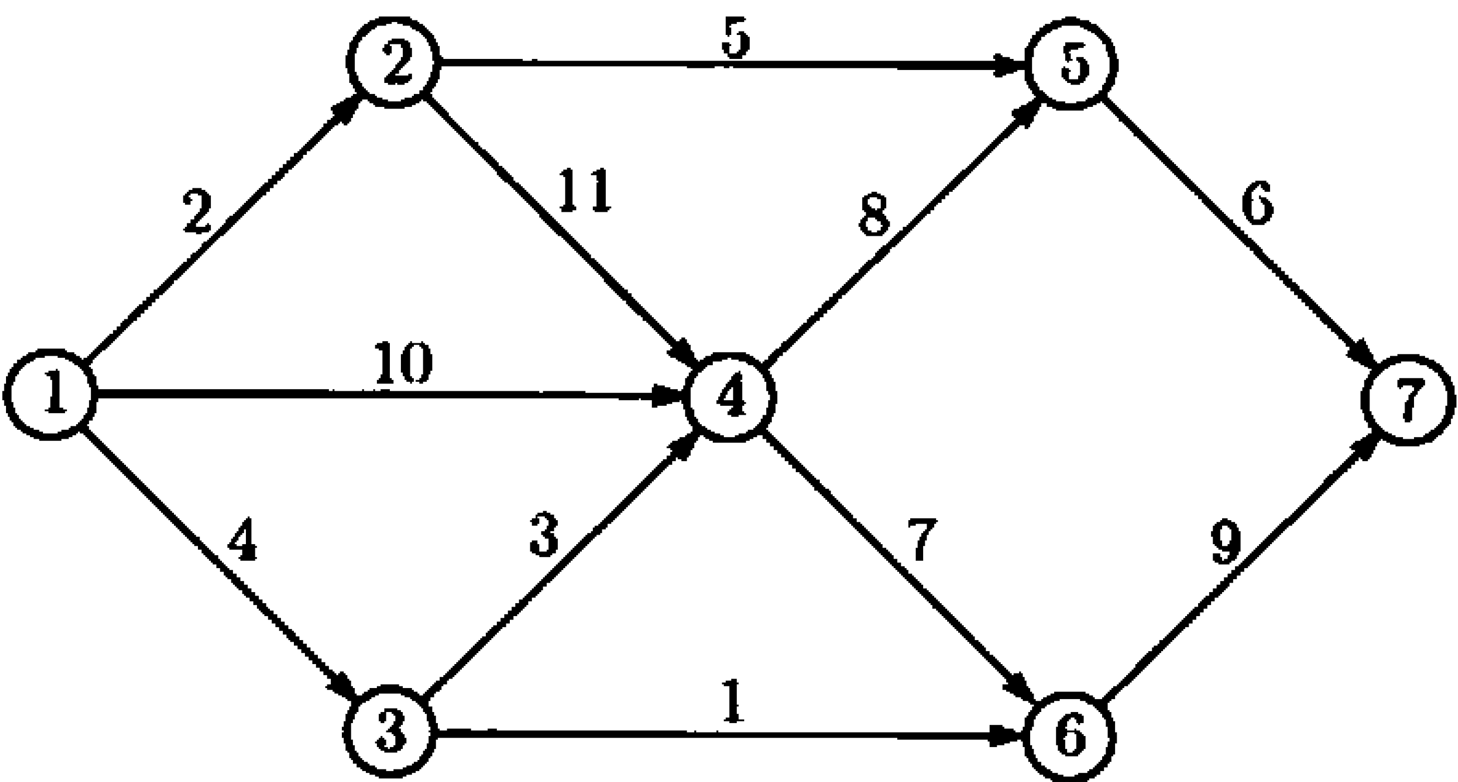


图 20.13 习题 20.1C 第 8 题的网络

9. 证明 6.4 节中的最大流模型是带有容量限制的最小费用流模型的一个特例, 并将这种转化方法应用到例 6.4-2 中. 为简便起见, 假定从节点 4 到节点 3 的容量为 0, 其他的数据保持不变.

10. 应用 Excel 规划求解和 AMPL 软件求解下面的问题:

(a) 习题 20.1A 第 3 题. (b) 习题 20.1A 第 4 题.

(c) 习题 20.1C 第 7 题. (d) 习题 20.1C 第 8 题.

20.2 分解算法

下面考虑为几台生产设备建立一套整体的方案, 虽然每台生产设备都有其自己的独立能力和生产的制约因素, 但是在公司考虑预算的层面上, 各种生产设备都连在了一起. 由此产生的模型包括两种类型的制约因素: *共同的*代表企业的预算约束, 而*独立的*则代表着每个设备的内部能力和生产限制. 图 20.14 详细描述了 n 个活动 (设备) 约束的布局, 在缺少共同约束的条件下所有的活动独立运行.

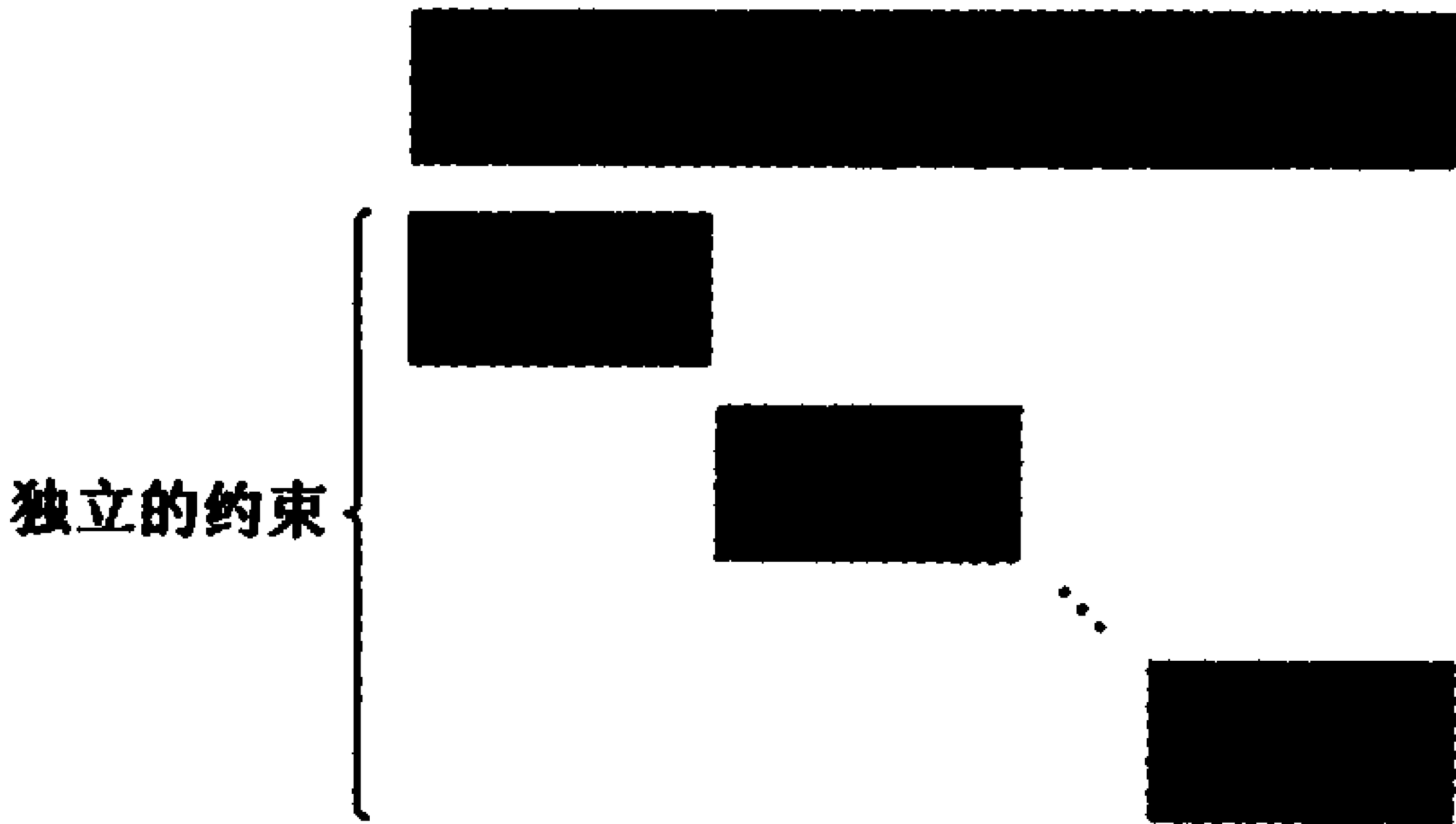


图 20.14 一个可分解的线性规划的结构

分解算法可以提高问题计算的效率, 例如对于图 20.14 描述的问题, 可以通过将这个问题分解成 n 个子问题, 然后独立地求解. 不过需要指出的是, 在以前计算机速度和内存比较适中的时候, 这种分解算法显得比较重要, 如今由于计算机性能的大幅提高, 使得我们忽略了使用这种分解算法. 然而我们之所以在此介绍这个算法, 是因为它的确是一个有趣的理论贡献.

下面给出相应的数学模型:

$$
\begin{array}{llllll}
\max & z = \boldsymbol{C}_1\boldsymbol{X}_1 + \boldsymbol{C}_2\boldsymbol{X}_2 + \cdots + \boldsymbol{C}_n\boldsymbol{X}_n & & & & \\
\text{s.t.} & \boldsymbol{A}_1\boldsymbol{X}_1 + & \boldsymbol{A}_2\boldsymbol{X}_2 + & \cdots + & \boldsymbol{A}_n\boldsymbol{X}_n & \leqslant \boldsymbol{b}_0 \\
& \boldsymbol{D}_1\boldsymbol{X}_1 & & & & \leqslant \boldsymbol{b}_1 \\
& & \boldsymbol{D}_2\boldsymbol{X}_2 & & & \leqslant \boldsymbol{b}_2 \\
& & \cdots\cdots & \cdots\cdots & & \\
& & & & \boldsymbol{D}_n\boldsymbol{X}_n & \leqslant \boldsymbol{b}_n \\
& \boldsymbol{X}_j \geqslant 0, j = 1, 2, \cdots, n & & & &
\end{array}
$$

通过添加松弛变量和剩余变量, 将所有的不等式约束转化为等式约束.

问题的分解原则是, 根据集合 $\{\boldsymbol{X}|\boldsymbol{D}_j\boldsymbol{X}_j \leqslant \boldsymbol{b}_j, \boldsymbol{X}_j \geqslant \boldsymbol{0}\}, j = 1, 2, \cdots, n$ 的极点来表示整个问题. 为了做到这一点, 要求每一个集合 $\{\boldsymbol{X}|\boldsymbol{D}_j\boldsymbol{X}_j \leqslant \boldsymbol{b}_j, \boldsymbol{X}_j \geqslant \boldsymbol{0}\}$ 的解空间都必须是有界的, 我们总可以通过对任意一个集合 j 增加人工约束 $\boldsymbol{1X}_j \leqslant M$ 来满足这个要求, 其中 M 是一个足够大的数.

用 $\hat{\boldsymbol{X}}_{jk}, k = 1, 2, \cdots, K_j$ 表示集合 $\{\boldsymbol{X}|\boldsymbol{D}_j\boldsymbol{X}_j \leqslant \boldsymbol{b}_j, \boldsymbol{X}_j \geqslant \boldsymbol{0}\}$ 的极点, 那么可以得到

$$
\boldsymbol{X}_j = \sum_{k=1}^{K_j} \beta_{jk}\hat{\boldsymbol{X}}_{jk}, \quad j = 1, 2, \cdots, n
$$

其中, 对所有的 k 都有 $\beta_{jk} \geqslant 0$, 并且 $\sum\limits_{k=1}^{K_j} \beta_{jk} = 1$.

根据极点集合, 可以将原来的整个问题重新建立成下面的主问题 (master problem) 模型:

$$
\begin{aligned}
\max\ z = & \sum_{k=1}^{K_1} \boldsymbol{C}_1\hat{\boldsymbol{X}}_{1k}\beta_{1k} + \sum_{k=1}^{K_2} \boldsymbol{C}_2\hat{\boldsymbol{X}}_{2k}\beta_{2k} + \cdots + \sum_{k=1}^{K_n} \boldsymbol{C}_n\hat{\boldsymbol{X}}_{nk}\beta_{nk} \\
\text{s.t.} & \sum_{k=1}^{K_1} \boldsymbol{A}_1\hat{\boldsymbol{X}}_{ik}\beta_{1k} + \sum_{k=1}^{K_2} \boldsymbol{A}_2\hat{\boldsymbol{X}}_{2k}\beta_{2k} + \cdots + \sum_{k=1}^{K_n} \boldsymbol{A}_n\hat{\boldsymbol{X}}_{nk}\beta_{nk} \leqslant \boldsymbol{b}_0 \\
& \sum_{k=1}^{K_j} \beta_{1k} = 1, \quad i = 1, 2, \cdots, n \\
& \beta_{jk} \geqslant 0, \text{ 对于所有的 } j \text{ 和 } k
\end{aligned}
$$

β_{jk} 是主问题中新的变量, 一旦找到了它们的最优解 β_{jk}^*, 那么就可以利用反向代换得到原始问题的最优解:

$$\boldsymbol{X}_j = \sum_{k=1}^{K_j} \beta_{jk}^* \hat{\boldsymbol{X}}_{jk}, \ j = 1, 2, \cdots, n$$

从这个解的形式上来看, 可能需要事先确定出所有的极点 $\hat{\boldsymbol{X}}_{jk}$ 才能求解主问题, 这当然是非常困难的一件事情, 但是幸运的是我们不需要这么做.

利用修正单纯形方法 (17.2 节) 求解主问题时, 需要确定每一次迭代中的进基和离基变量. 对于如何确定进基变量, 假定给出了主问题当前基的 $\boldsymbol{C}_B$ 和 $\boldsymbol{B}^{-1}$, 那么对于非基变量 β_{jk}, 我们有

$$z_{jk} - c_{jk} = \boldsymbol{C}_B \boldsymbol{B}^{-1} \boldsymbol{P}_{jk} - c_{jk}$$

其中

$$c_{jk} = \boldsymbol{C}_j \hat{\boldsymbol{X}}_{jk}, \ \boldsymbol{P}_{jk} = \begin{pmatrix} \boldsymbol{A}_j \hat{\boldsymbol{X}}_{jk} \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$$

如果存在多个可以进基的非基变量 β_{jk}, 且要确定到底哪一个进基, 那么就需要确定

$$z_{j^*k^*} - c_{j^*k^*} = \min_{\text{所有的 } j \text{ 和 } k} \{z_{jk} - c_{jk}\}$$

如果 $z_{j^*k^*} - c_{j^*k^*} > 0$, 那么根据最大最优性条件, 变量 $\beta_{j^*k^*}$ 进基, 否则已经达到了最优.

目前我们仍然没有给出如何计算 $z_{j^*k^*} - c_{j^*k^*}$ 的值, 需要根据下面的式子计算:

$$\min_{\text{所有的 } j \text{ 和 } k} \{z_{jk} - c_{jk}\} = \min_j \{\min_k \{z_{jk} - c_{jk}\}\}$$

之所以有上面的式子是因为, 每一个凸集合 $\{\boldsymbol{X} | \boldsymbol{D}_j \boldsymbol{X}_j \leqslant \boldsymbol{b}_j, \boldsymbol{X}_j \geqslant \boldsymbol{0}\}$ 都有自己独立的极点集合. 因此, 根据上面式子给出的结论, 可以按照下面的两个步骤求得 $z_{j^*k^*} - c_{j^*k^*}$:

第 1 步 对于每一个凸集合 $\{\boldsymbol{X} | \boldsymbol{D}_j \boldsymbol{X}_j \leqslant \boldsymbol{b}_j, \ \boldsymbol{X}_j \geqslant \boldsymbol{0}\}$, 确定使得 $z_{jk} - c_{jk}$ 达到最小的极点 $\hat{\boldsymbol{Y}}_{jk^*}$, 即 $z_{jk^*} - c_{jk^*} = \min_k \{z_{jk} - c_{jk}\}$.

第 2 步　确定 $z_{j^*k^*} - c_{j^*k^*} = \min\limits_{j}\{z_{jk^*} - c_{jk^*}\}$.

根据线性规划理论, 我们知道当问题的最优解存在并且最优值有限的时候, 那么一定在解空间的一个极点上达到最优. 因为根据定义可以知道, 每一个集合 $\{\boldsymbol{X}|\boldsymbol{D}_j\boldsymbol{X}_j \leqslant \boldsymbol{b}_j, \boldsymbol{X}_j \geqslant \boldsymbol{0}\}$ 都是有界的, 所以第一步等价于求解 n 个线性规划

$$\min\ w_j = \{z_j - c_j|\boldsymbol{D}_j\boldsymbol{X}_j \leqslant \boldsymbol{b}_j,\quad \boldsymbol{X}_j \geqslant \boldsymbol{0}\}$$

目标函数 w_j 是关于 $\boldsymbol{X}_j$ 的一个线性函数 (见习题 20.2A 第 8 题).

在主问题中, 确定进基变量 $\beta_{j^*k^*}$ 的过程, 简化为求解 n 个较小的线性规划以确定进基的极点 $\hat{\boldsymbol{Y}}_{j^*k^*}$. 利用这种方法就避免了确定所有 n 个凸集合中的所有极点, 一旦得到了所需要的极点, 那么这个列向量 $\boldsymbol{P}_{j^*k^*}$ 中所有的元素就都可以得到, 然后就可以确定离开基的变量, 因此也就可以利用修正的单纯形算法得到下一个基 $\boldsymbol{B}^{-1}$.

例 20.2-1

利用分解算法求解下面的线性规划:

$$\begin{aligned}
\max\quad & z = 3x_1 + 5x_2 + x_3 + x_4 \\
\text{s.t.}\quad & x_1 + x_2 + x_3 + x_4 \leqslant 40 \\
& 5x_1 + x_2 \leqslant 12 \\
& x_3 + x_4 \geqslant 5 \\
& x_3 + 5x_4 \leqslant 50 \\
& x_1, x_2, x_3, x_4 \geqslant 0
\end{aligned}$$

根据下面变量的分解方法可以得到两个子问题:

$$\boldsymbol{X}_1 = (x_1, x_2)^{\mathrm{T}}, \qquad \boldsymbol{X}_2 = (x_3, x_4)^{\mathrm{T}}$$

于是, 相应的主问题是

通过增加松弛变量 x_5 将原来的不等式约束转化为下面的等式约束:

$$x_1 + x_2 + x_3 + x_4 + x_5 = 40$$

又注意到子问题 1 和子问题 2 只与变量 x_1, x_2, x_3, x_4 有关, 这就是为什么主问题中必须明确地包含变量 x_5. 初始基本解中的另外两个变量 x_6 和 x_7 是人工变量.

子问题 1				子问题 2				初始基本解			
β_{11}	β_{12}	$\cdots$	β_{1K_1}	β_{21}	β_{22}	$\cdots$	β_{2K_2}	x_5	x_6	x_7	
$C_1\hat{X}_{11}$	$C_1\hat{X}_{12}$	$\cdots$	$C_1\hat{X}_{1K_1}$	$C_2\hat{X}_{21}$	$C_2\hat{X}_{22}$	$\cdots$	$C_2\hat{X}_{2K_2}$	0	$-M$	$-M$	
$A_1\hat{X}_{11}$	$A_1\hat{X}_{12}$	$\cdots$	$A_1\hat{X}_{1K_1}$	$A_2\hat{X}_{21}$	$A_2\hat{X}_{22}$	$\cdots$	$A_2\hat{X}_{2K_2}$	1	0	0	= 40
1	1	$\cdots$	1	0	0	$\cdots$	0	0	1	0	= 1
0	0	$\cdots$	0	1	1	$\cdots$	1	0	0	1	= 1

$C_1 = (3,5)$　　　　$C_2 = (1,1)$

$A_1 = (1,1)$　　　　$A_2 = (1,1)$

解空间, $D_1X_1 \leqslant b_1$:

$$5x_1 + x_2 \leqslant 12$$
$$x_1, x_2 \geqslant 0$$

解空间, $D_2X_2 \leqslant b_2$:

$$x_3 + x_4 \geqslant 5$$
$$x_3 + 5x_4 \leqslant 50$$
$$x_3, x_4 \geqslant 0$$

迭代 0

$$X_B = (x_5, x_6, x_7)^{\mathrm{T}} = (40, 1, 1)^{\mathrm{T}}$$
$$C_B = (0, -M, -M), \quad B = B^{-1} = I$$

迭代 1

子问题 $1(j=1)$. 我们有

$$\begin{aligned} z_1 - c_1 &= C_B B^{-1} \begin{pmatrix} A_1X_1 \\ 1 \\ 0 \end{pmatrix} - C_1X_1 \\ &= (0, -M, -M) \begin{pmatrix} (1,1)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ 1 \\ 0 \end{pmatrix} - (3,5)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= -3x_1 - 5x_2 - M \end{aligned}$$

那么相应的线性规划问题是:

$$\begin{aligned} \min\ & w_1 = -3x_1 - 5x_2 - M \\ \text{s.t.}\ & 5x_1 + x_2 \leqslant 12 \\ & x_1,\ x_2 \geqslant 0 \end{aligned}$$

通过单纯形算法可以得到这个问题的解:

$$\hat{X}_{11} = (0, 12)^{\mathrm{T}}, \quad z_1^* - c_1^* = w_1^* = -60 - M$$

子问题 $2(j=2)$. 相应的线性规划是:

$$
\begin{aligned}
\min\ z_2 - c_2 &= \boldsymbol{C}_B\boldsymbol{B}^{-1}\begin{pmatrix}\boldsymbol{A}_2\boldsymbol{X}_2\\0\\1\end{pmatrix} - \boldsymbol{C}_2\boldsymbol{X}_2\\
&= (0,-M,-M)\begin{pmatrix}(1,1)\begin{pmatrix}x_3\\x_4\end{pmatrix}\\0\\1\end{pmatrix} - (1,1)\begin{pmatrix}x_4\\x_5\end{pmatrix}\\
&= -x_4 - x_5 - M\\
\text{s.t. } x_3 + x_4 &\geqslant 5\\
x_3 + 5x_4 &\leqslant 50\\
x_3,\ x_4 &\geqslant 0
\end{aligned}
$$

这个问题的最优解是:

$$\hat{\boldsymbol{X}}_{21} = (50,0)^{\mathrm{T}},\quad z_2^* - c_2^* = -50 - M$$

因为主问题的目标是求最大, 并且 $z_1^* - c_1^* < z_2^* - c_2^*$ 和 $z_1^* - c_1^* < 0$, 所以可以得到相应极点 $\hat{\boldsymbol{X}}_{11}$ 的 β_{11} 进基. 下面确定离开基的变量.

$$\boldsymbol{P}_{11} = \begin{pmatrix}\boldsymbol{A}_1\hat{\boldsymbol{X}}_{11}\\1\\0\end{pmatrix} = \begin{pmatrix}(1,1)\begin{pmatrix}0\\12\end{pmatrix}\\1\\0\end{pmatrix} = \begin{pmatrix}12\\1\\0\end{pmatrix}$$

因此 $\boldsymbol{B}^{-1}\boldsymbol{P}_{11} = (12,1,0)^{\mathrm{T}}$. 给定 $\boldsymbol{X}_B = (x_5,x_6,x_7)^{\mathrm{T}} = (40,1,1)^{\mathrm{T}}$, 可以知道人工变量 x_6 是 (永久的) 离开基本解的变量.

用向量 $\boldsymbol{P}_{11}$ 来代替变量 x_6 对应的向量, 可以得到新的基 (证明!)

$$\boldsymbol{B} = \begin{pmatrix}1 & 12 & 0\\0 & 1 & 0\\0 & 0 & 1\end{pmatrix} \Rightarrow \boldsymbol{B}^{-1} = \begin{pmatrix}1 & -12 & 0\\0 & 1 & 0\\0 & 0 & 1\end{pmatrix}$$

新的基本解是:

$$
\begin{aligned}
\boldsymbol{X}_B &= (x_5,\beta_{11},x_7)^{\mathrm{T}} = \boldsymbol{B}^{-1}(40,1,1)^{\mathrm{T}} = (28,1,1)^{\mathrm{T}}\\
\boldsymbol{C}_B &= (0,\boldsymbol{C}_{1,11},-M) = (0,60,-M)
\end{aligned}
$$

迭代 2

子问题 $1(j=1)$. 对应的目标函数是:

$$\min\ w_1 = -3x_1 - 5x_2 + 60$$

(请证明!). 由最优解可以得到 $z_1^* - c_1^* = w_1 = 0$, 所以子问题 1 中不再有极点可以改进主问题的解了.

子问题 2($j=2$). 对应的目标函数与迭代 1 中 $j=2$ 对应的目标函数相同 (请证明!), 可以得到最优解是:

$$\hat{\boldsymbol{X}}_{22}=(50,0)^{\mathrm{T}},\quad z_2^*-c_2^*=-50-M$$

注意到 $\hat{\boldsymbol{X}}_{22}$ 与 $\hat{\boldsymbol{X}}_{21}$ 相同, 下标的 2 是为了方便表示第 2 次迭代.

根据这两个子问题的解, $z_2^*-c_2^*<0$ 表示相应于 $\hat{\boldsymbol{X}}_{22}$ 的 β_{22} 进基.

下面确定离开基的变量, 考虑

$$\boldsymbol{P}_{22}=\begin{pmatrix}\boldsymbol{A}_2\hat{\boldsymbol{X}}_{22}\\0\\1\end{pmatrix}=\begin{pmatrix}(1,1)\begin{pmatrix}50\\0\end{pmatrix}\\0\\1\end{pmatrix}=\begin{pmatrix}50\\0\\1\end{pmatrix}$$

故有 $\boldsymbol{B}^{-1}\boldsymbol{P}_{22}=(50,0,1)^{\mathrm{T}}$. 因为 $\boldsymbol{X}_B=(x_5,\beta_{11},x_7)^{\mathrm{T}}=(28,1,1)^{\mathrm{T}}$, 所以 x_5 离开基.

新的基和它的逆是 (请证明!):

$$\boldsymbol{B}=\begin{pmatrix}50&12&0\\0&1&0\\1&0&1\end{pmatrix}\Rightarrow\boldsymbol{B}^{-1}=\begin{pmatrix}\frac{1}{50}&-\frac{12}{50}&0\\0&1&0\\-\frac{1}{50}&\frac{12}{50}&1\end{pmatrix}$$

$$\boldsymbol{X}_B=(\beta_{22},\beta_{11},x_7)^{\mathrm{T}}=\boldsymbol{B}^{-1}(40,1,1)^{\mathrm{T}}=\left(\frac{14}{25},1,\frac{11}{25}\right)^{\mathrm{T}}$$

$$\boldsymbol{C}_B=(\boldsymbol{C}_{2,22},\boldsymbol{C}_{1,11},-M)=(50,60,-M)$$

迭代 3

子问题 1($j=1$). 需要证明相应的目标函数是:

$$\min\ w_1=\left(\frac{M}{50}-2\right)x_1+\left(\frac{M}{50}-4\right)x_2-\frac{12M}{50}+48$$

相应的最优解是:

$$\hat{\boldsymbol{X}}_{13}=(0,0)^{\mathrm{T}},\quad z_1^*-c_1^*=-\frac{12M}{50}+48$$

子问题 2($j=2$). 目标函数为 (证明!):

$$\min\ w_2=\left(\frac{M}{50}\right)(x_3+x_4)-M$$

相应的最优解是:

$$\hat{\boldsymbol{X}}_{23}=(5,0)^{\mathrm{T}},\quad z_2^*-c_2^*=-\frac{9M}{10}$$

非基变量 x_5. 根据主问题的定义, 计算对应于 x_5 的 z_j-c_j 并比较, 得到

$$\begin{aligned}z_5-c_5&=\boldsymbol{C}_B\boldsymbol{B}^{-1}\boldsymbol{P}_5-c_5\\&=\left(1+\frac{M}{50},48-\frac{12M}{50},-M\right)(1,0,0)^{\mathrm{T}}-0=1+\frac{M}{50}\end{aligned}$$

因此, x_5 不能改进问题的解.

根据上面的信息可知, 对应于 $\hat{X}_{23}$ 的 β_{23} 进入基. 下面需要确定离开基的变量,

$$P_{23}=\begin{pmatrix} A_2\hat{X}_{23} \\ 0 \\ 1 \end{pmatrix}=\begin{pmatrix} (1,1)\begin{pmatrix}5\\0\end{pmatrix} \\ 0 \\ 1 \end{pmatrix}=\begin{pmatrix}5\\0\\1\end{pmatrix}$$

故有 $B^{-1}P_{23}=\left(\frac{1}{10},0,\frac{9}{10}\right)^{\mathrm{T}}$. 给定 $X_B=(\beta_{22},\beta_{11},x_7)^{\mathrm{T}}=\left(\frac{14}{25},1,\frac{11}{25}\right)^{\mathrm{T}}$, 可以知道人工变量 x_{10} 是 (永久的) 离开基本解的变量.

新的基和它的逆是 (请证明!):

$$B=\begin{pmatrix} 50 & 12 & 5 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \Rightarrow B^{-1}=\begin{pmatrix} \frac{1}{45} & -\frac{12}{45} & -\frac{5}{45} \\ 0 & 1 & 0 \\ -\frac{1}{45} & \frac{12}{45} & \frac{50}{45} \end{pmatrix}$$

$$X_B=(\beta_{22},\beta_{11},\beta_{23})^{\mathrm{T}}=B^{-1}(40,1,1)^{\mathrm{T}}=\left(\frac{23}{45},1,\frac{22}{45}\right)^{\mathrm{T}}$$

$$C_B=(C_{2,22},C_{1,11},C_{2,23})=(50,60,5)$$

迭代 4

子问题 $1(j=1)$. $w_1=-2x_1-4x_2+48$, 可知 $z_1^*-c_1^*=w_1^*=0$.

子问题 $2(j=2)$. $w_2=0x_3+0x_4$, $w_2^*=0$.

非基变量 x_5: $z_5-c_5=1$. 根据上面的信息可以知道迭代 3 是最优的.

通过变量的反向代换可以计算出原始问题的最优解是:

$$X_1^*=(x_1,x_2)^{\mathrm{T}}=\beta_{11,11}=1(0,12)^{\mathrm{T}}=(0,12)^{\mathrm{T}}$$

$$X_2^*=(x_4,x_5)^{\mathrm{T}}=\beta_{22,22}+\beta_{23,23}$$

$$=\left(\frac{23}{45}\right)(50,0)^{\mathrm{T}}+\left(\frac{22}{45}\right)(5,0)^{\mathrm{T}}=(28,0)^{\mathrm{T}}$$

剩余的其他变量都等于 0. 目标函数的最优值可以通过直接变量代换求出.

习题 20.2A

1. 利用图形确定下面每一个例子的可行极点集合, 并且将可行解空间表达成这些极点的函数. 如果解空间是无界的, 那么添加恰当的人工约束.

(a) $x_1+2x_2\leqslant 6$
$2x_1+x_2\leqslant 8$
$-x_1+x_2\leqslant 1$
$x_2\leqslant 2$
$x_1,x_2\leqslant 0$

(b) $2x_1+x_2\leqslant 2$
$3x_1+4x_2\geqslant 12$
$x_1,x_2\geqslant 0$

*(c) $x_1-x_2\leqslant 10$
$2x_1\leqslant 40$
$x_1,x_2\geqslant 0$

*2. 在例 20.2-1 中，子空间 $D_1X_1 = b_1$，$X_1 \geqslant 0$ 和 $D_2X_2 = b_2$，$X_2 \geqslant 0$ 的极点集合可以用图示法得到．利用这一信息，根据可行极点集合准确地表示相应的主问题．然后证明：应用单纯形算法求解主问题产生的进基变量 β_{jk}，与分别求解子问题 1 和子问题 2 得到的相同．因此，可以确信这样得到的进基变量 β_{jk} 恰好等价于求解两个最小化子问题．

3. 考虑下面的线性规划：

$$
\begin{aligned}
\max\ z = x_1 + 3x_2 + 5x_3 + 2x_4 & \\
\text{s.t.}\ x_1 + 4x_2 & \leqslant 8 \\
2x_1 + x_2 & \leqslant 9 \\
5x_1 + 3x_2 + 4x_3 & \geqslant 10 \\
x_3 - 5x_4 & \leqslant 4 \\
x_3 + x_4 & \leqslant 10 \\
x_1,\ x_2,\ x_3,\ x_4 & \geqslant 0
\end{aligned}
$$

使用子空间的极点集合，建立一个准确的主问题，并且采用单纯形算法求解得到的主问题．

4. 利用分解算法求解第 3 题，并将这两种求解过程进行比较．

5. 利用分解算法求解下面的问题：

$$
\begin{aligned}
\max\ z = 6x_1 + 7x_2 + 3x_3 + 5x_4 + x_5 + x_6 & \\
\text{s.t.}\ x_1 + x_2 + x_3 + x_4 + x_5 + x_6 & \leqslant 50 \\
x_1 + x_2 & \leqslant 10 \\
x_2 & \leqslant 8 \\
5x_3 + x_4 & \leqslant 12 \\
x_5 + x_6 & \geqslant 5 \\
x_5 + 5x_6 & \leqslant 50 \\
x_1, x_2, x_3, x_4, x_5, x_6 & \geqslant 0
\end{aligned}
$$

*6. 指明在利用分解算法求解最小化线性规划时需要进行的必要转化．然后求解下面的问题：

$$
\begin{aligned}
\min\ z = 5x_1 + 3x_2 + 8x_3 - 5x_4 & \\
\text{s.t.}\ x_1 + x_2 + x_3 + x_4 & \geqslant 25 \\
5x_1 + x_2 & \leqslant 20 \\
5x_1 - x_2 & \geqslant 5 \\
x_3 + x_4 & = 20 \\
x_1, x_2, x_3, x_4 & \geqslant 0
\end{aligned}
$$

7. 利用分解算法求解下面的问题：

$$
\begin{aligned}
\min\ z = 10y_1 + 2y_2 + 4y_3 + 8y_4 + y_5 & \\
\text{s.t.}\ y_1 + 4y_2 - y_3 & \geqslant 8 \\
2y_1 + y_2 + y_3 & \geqslant 2 \\
3y_1 + y_4 + y_5 & \geqslant 4 \\
y_1 + 2y_4 - y_5 & \geqslant 10 \\
y_1, y_2, y_3, y_4, y_5 & \geqslant 0
\end{aligned}
$$

(提示: 首先利用分解算法求解对偶问题.)

8. 在分解算法中, 假定原始问题中公共约束的数目是 r, 证明子问题 j 的目标函数可以写作:

$$\min\ w_j = z_j - c_j = (\boldsymbol{C}_B\boldsymbol{R}\boldsymbol{A}_j - \boldsymbol{C}_j)\boldsymbol{X}_j + \boldsymbol{C}_B\boldsymbol{V}_{r+j}$$

向量 $\boldsymbol{R}$ 表示 $\boldsymbol{B}^{-1}$ 的前 r 列, 向量 $\boldsymbol{V}_{r+j}$ 是 $\boldsymbol{B}^{-1}$ 的第 $r+j$ 列.

20.3 Karmarkar 内点算法

单纯形方法是沿着解空间的边界上一条由紧邻的极点组成的路径到达最优解的. 虽然实际中单纯形方法也可以很有效地求解大规模问题, 但是理论上找到最优解的迭代次数可能会是指数级别的. 实际上, 研究者已经构造出了一类线性规划问题, 当利用单纯形求解这类问题时会访问到所有的极点才能找到最优解.

1984 年, N. Karmarkar 设计出了一个通过切割解空间求解线性规划的多项式时间算法, 这种算法对于求解大规模线性规划尤其有效.

我们首先介绍 Karmarkar 算法的主要思想, 然后采用例题详细介绍它.

20.3.1 内点算法的基本思想

考虑下面简单的例题:

$$\max\ z = x_1$$
$$\text{s.t.}\ 0 \leqslant x_1 \leqslant 2$$

引入松弛变量 x_2, 这个问题可以重新记作:

$$\max\ z = x_1$$
$$\text{s.t.}\ x_1 + x_2 = 2$$
$$x_1, x_2 \geqslant 0$$

图 20.15 描述了这个问题, 解空间是线段 AB, x_1 的正向是 z 增加的方向.

下面我们任意选取可行解空间 (直线 AB) 内的一个点 C(非极点), 目标函数 ($\max z = x_1$) 在 C 点的**梯度** (gradient) 是指 z 增长最快的方向. 如果我们在梯度的方向上任意取一点, 然后再将这个点垂直地映射到可行解区域 (直线 AB) 上, 那么就可以得到一个使得目标函数 z 更优的新点 D, 这样的一个改进过程可以通过沿着映射后的梯度 CD 的方向得到. 在 D 点重复上面的过程, 可以得到一个更加靠近最优解的点 E. 可以预料到, 如果我们 (谨慎地) 沿着**映射梯度** (projected gradient) 的方向移动, 最终将达到最优解 B 点. 当我们最小化目标函数 z 的时候, 映射梯度恰好引导我们远离 B 而转向最小值点 $A(x_1 = 0)$.

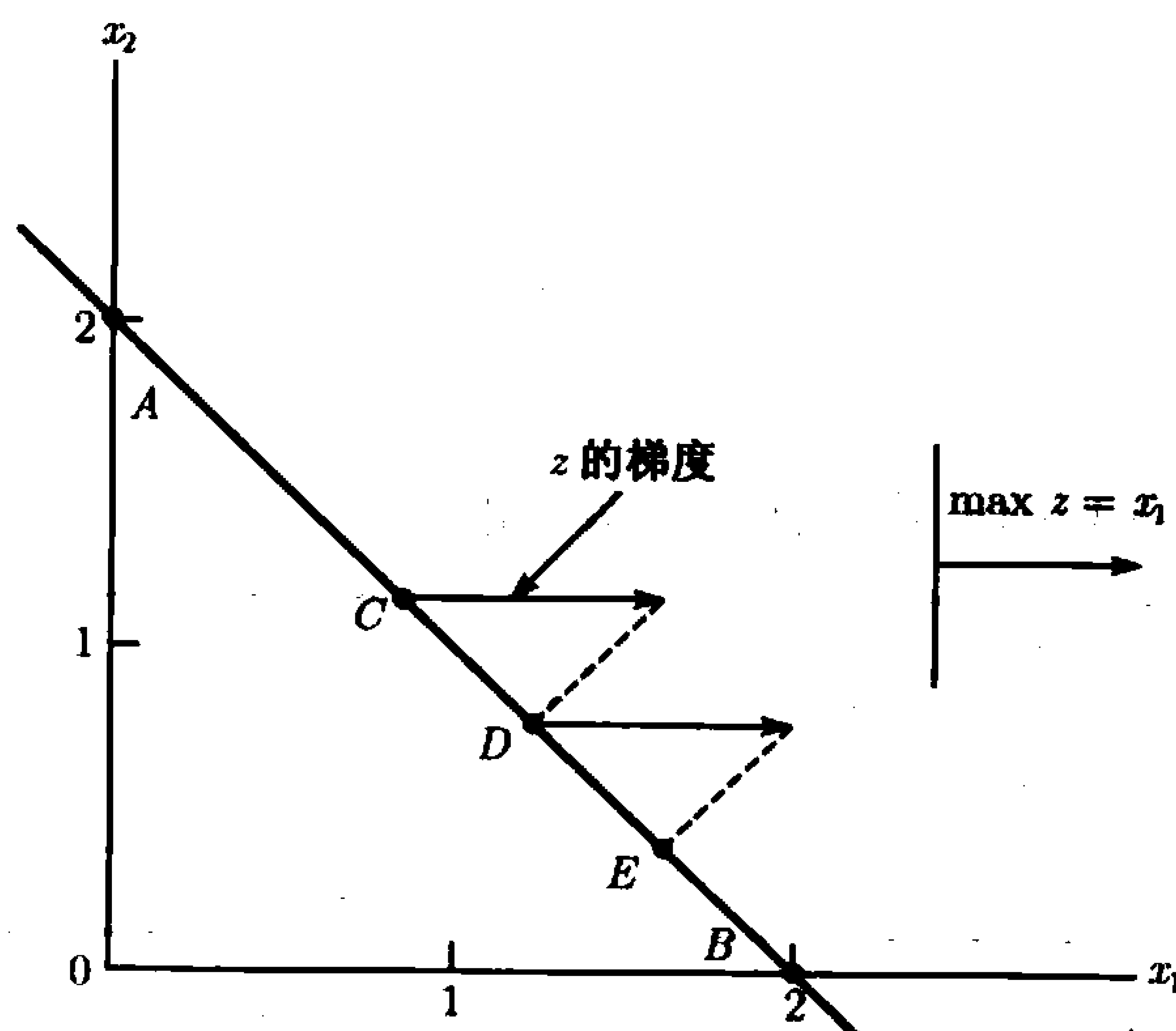


图 20.15 Karmarkar 算法的一般思想

上面的过程给出了算法的一般概念, 但是思想还是有些模糊! 我们需要某些修正来保证 (1) 沿着映射梯度的迭代不会“跳过”最优解点 B, (2) 在一般的 n 维空间的情况下, 沿着映射梯度给出的方向迭代时, 算法不会局限在一个非最优的点周围. 这两点也正是 Karmarkar 内点算法所需要完成的基本任务.

20.3.2 内点算法

在不同的文献中, Karmarkar 算法的描述也有所不同. 我们将要介绍的是初始的算法. Karmarkar 假定一个线性规划问题有下面的形式:

$$
\begin{aligned}
\min \quad & z = \boldsymbol{CX} \\
\text{s.t.} \quad & \boldsymbol{AX} = \boldsymbol{0} \\
& \boldsymbol{1X} = 1 \\
& \boldsymbol{X} \geqslant \boldsymbol{0}
\end{aligned}
$$

除了约束 $\boldsymbol{1X} = \sum\limits_{j=1}^{n} x_j = 1$ 定义了一个 n 维单纯形之外, 其他的约束都是齐次的等式. Karmarkar 算法的合理性是建立在下面两个条件基础之上的:

(1) $\boldsymbol{X} = \left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right)$ 满足 $\boldsymbol{AX} = \boldsymbol{0}$;

(2) $\min z = 0$.

Karmarkar 也提供了修正的方法来求解不满足第 (2) 个条件的问题, 这里就不再介绍具体的修正方法了.

通过下面的例题, 我们介绍如何得到一个一般的线性规划以满足上面的两个条件.

例 20.3-1

考虑下面的问题:

$$\begin{aligned} \min\ & z = y_1 + y_2 \\ \text{s.t.}\ & y_1 + 2y_2 \leqslant 2 \\ & y_1,\ y_2 \geqslant 0 \end{aligned}$$

首先写出这个线性规划的原始问题和对偶问题:

原	始	对	偶
max	$y_0 = y_1 + y_2$	min	$w_0 = 2w_1$
s.t.	$y_1 + 2y_2 \leqslant 2$	s.t.	$\left.\begin{aligned} w_1 &\geqslant 1 \\ 2w_1 &\geqslant 1 \end{aligned}\right\} \Rightarrow w_1 \geqslant 1$
	$y_1, y_2 \geqslant 0$		$w_1 \geqslant 0$

原始问题和对偶问题的约束可以写成下面的等式形式:

$$\begin{aligned} y_1 + 2y_2 + y_3 = 2, &\quad y_3 \geqslant 0 \\ w_1 - w_2 = 1, &\quad w_2 \geqslant 0 \end{aligned} \tag{20.1}$$

最优解满足 $y_0 = w_0$, 所以

$$y_1 + y_2 - 2w_1 = 0 \tag{20.2}$$

设 M 是一个足够大的数, 则

$$y_1 + y_2 + y_3 + w_1 + w_2 \leqslant M \tag{20.3}$$

把上面的不等式转化为等式, 得到

$$y_1 + y_2 + y_3 + w_1 + w_2 + s_1 = M, s_1 \geqslant 0 \tag{20.4}$$

下面定义一个新的变量 s_2. 当且仅当 $s_2 = 1$ 时, 根据 (20.4) 可以得到下面的两个等式:

$$\begin{aligned} y_1 + y_2 + y_3 + w_1 + w_2 + s_1 - Ms_2 &= 0 \\ y_1 + y_2 + y_3 + w_1 + w_2 + s_1 + \quad s_2 &= M + 1 \end{aligned} \tag{20.5}$$

现在给定 $s_2 = 1$, 那么原始对偶等式 (20.1) 可以写成:

$$\begin{aligned} y_1 + 2y_2 + y_3 - 2s_2 &= 0 \\ w_1 - w_2 - 1s_2 &= 0 \end{aligned} \tag{20.6}$$

定义

$$\begin{aligned} y_j &= (M+1)x_j, \quad j=1,2,3 \\ w_{j-3} &= (M+1)x_j, \quad j=4,5 \\ s_1 &= (M+1)x_6 \\ s_2 &= (M+1)x_7 \end{aligned}$$

对等式 (20.2),(20.5) 和 (20.6) 进行变量代换, 可以得到下面的等式组:

$$\begin{aligned} &x_1 + x_2 - 2x_4 = 0 \\ &x_1 + x_2 + x_3 + x_4 + x_5 + x_6 - Mx_7 = 0 \\ &x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 = 1 \\ &x_1 + 2x_2 + x_3 - 2x_7 = 0 \\ &x_4 - x_5 - x_7 = 0 \\ &x_j \geqslant 0, \quad j = 1,2,\cdots,7 \end{aligned}$$

最后一步就是要在每一个等式左边增加人工变量 y_8, 新的目标函数将最小化 y_8, 而这个变量明显的最小值是 0(假定原始问题有可行解的情况下). 然而, 注意到 Karmarkar 算法要求解

$$\boldsymbol{X} = \left(\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right)^{\mathrm{T}}$$

满足方程组 $\boldsymbol{AX} = \boldsymbol{0}$, 当所有齐次方程 (右边为 0) 中对应于人工变量 x_8 的系数等于左边所有系数的和的时候, 上面的条件就是满足的. 下面给出转化了的线性规划:

$$\begin{aligned} \min\ & z = x_8 \\ \text{s.t. } & x_1 + x_2 - 2x_4 - 0x_8 = 0 \\ & x_1 + x_2 + x_3 + x_4 + x_5 + x_6 - Mx_7 - (6-M)x_8 = 0 \\ & x_1 + 2x_2 + x_3 - 2x_7 - 2x_8 = 0 \\ & x_4 - x_5 - x_7 + x_8 = 0 \\ & x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 = 1 \\ & x_j \geqslant 0, \quad j = 1,2,\cdots,8 \end{aligned}$$

利用变量反向代换, 可以由上面的解自动得到原始对偶问题的最优解.

图 20.16 提供了一个典型的三维空间上解空间的描述, 其中只包含一个等式约束. 根据定义, 解空间是由满足二维单纯形 $\boldsymbol{1X} = 1$ 上的一条线段 AB 组成的, 并且 AB 通过内部可行点 $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

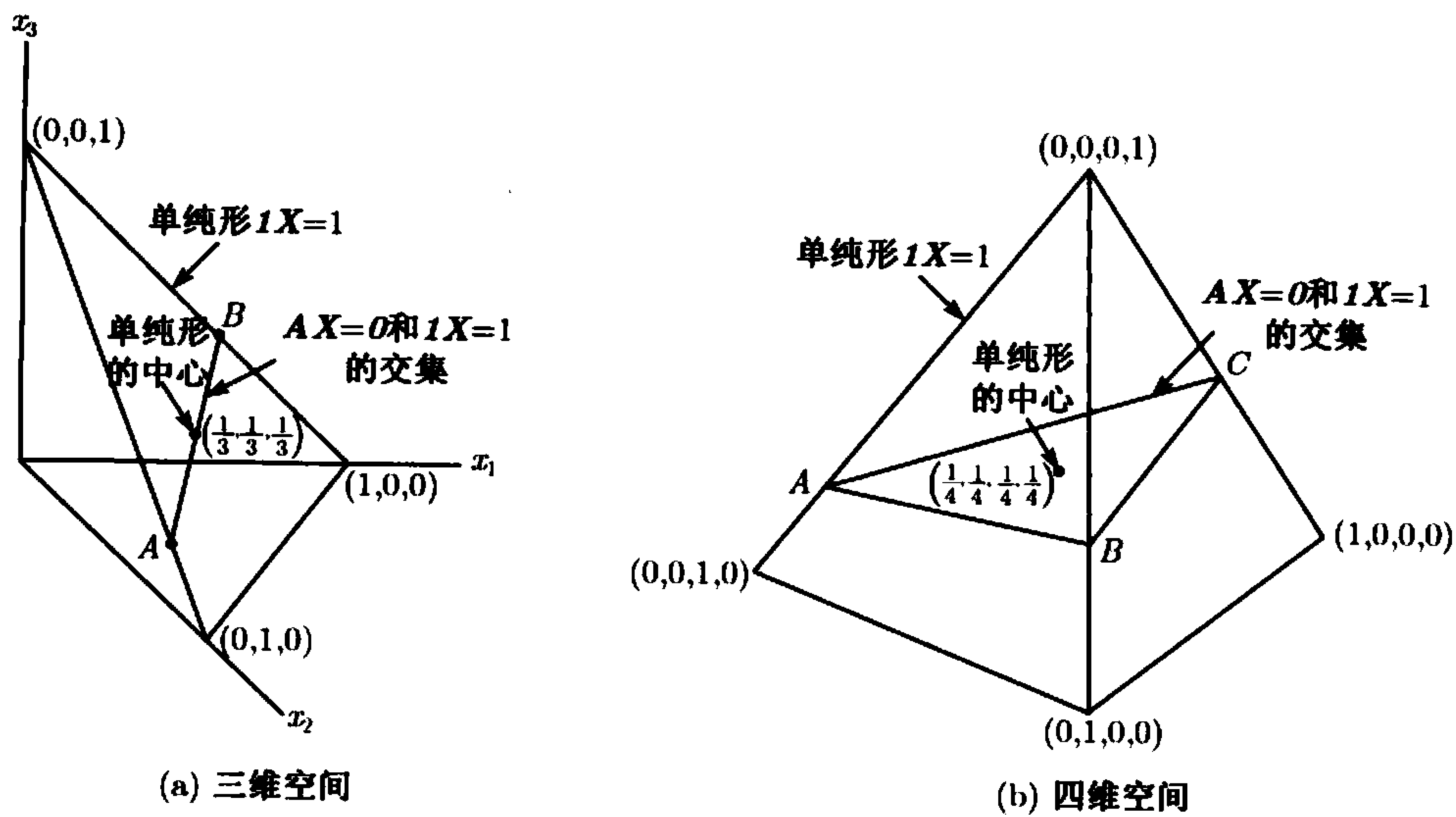


图 20.16　单纯形 $\mathit{1}X = 1$ 的图示

算法步骤　Karmarkar 算法从表示单纯形中心的内点出发, 然后在梯度映射方向上移动以得到下一个新的可行点. 新的可行点一定是严格地在可行区域内部, 也就是它的所有坐标值都是正数. 算法的可行性就是建立在这个条件基础上的.

因为新的可行解点是一严格内点, 它一定不在单纯形的边界上. (对于图 20.16, 点 A 和 B 一定要排除掉.) 为了保证这一点, 必须使得内点有一个球形邻域, 这个邻域完全落在单纯形内部. 在 n 维空间的情形下, 这个球的半径等于 $\frac{1}{\sqrt{n(n-1)}}$, 同时半径为 $\alpha r(0 < \alpha < 1)$ 的球一定是这个邻域的子集, 并且在这个球与齐次系统 $AX = \mathit{0}$ 的交集中的任意一点都是一个内点, 坐标都是正数. 这样我们就可以在这个限制的区域内 ($AX = \mathit{0}$ 和 αr−球的交集), 沿着梯度映射的方向尽可能远地找到一个新的 (改进的) 可行解.

新的可行解不再位于单纯形的中心了. 为了使这个过程可以进行迭代, 我们需要找到一种方法使得新的可行解成为一个单纯形的中心. Karmarkar 提出了解决这个问题的一个大概的思想, 称作**射影变换**(projective transformation). 令

$$y_i = \frac{\frac{x_i}{x_{ki}}}{\sum_{j=1}^{n} \frac{x_j}{x_{kj}}}, \quad i = 1, 2, \cdots, n$$

其中 x_{ki} 表示当前可行点 X_k 的第 i 个分量. 由于所有的 $x_{ki} > 0$, 所以变换是合理的, 并且根据定义有 $\sum_{i=1}^{n} y_i = 1$ 或者 $\mathit{1}Y = 1$. 这个变换也就等价于

$$Y = \frac{D_k^{-1}X}{\mathit{1}D_k^{-1}X}$$

其中 $\boldsymbol{D}_k$ 是对角矩阵, 第 i 个对角元素等于 x_{ki}. 因为我们根据上面的等式可以得到下面的式子, 所以这个变换是将 X−空间唯一地映射到 Y−空间上:

$$\boldsymbol{X} = \frac{\boldsymbol{D}_k\boldsymbol{Y}}{\boldsymbol{1}\boldsymbol{D}_k\boldsymbol{Y}}$$

根据定义, $\min \boldsymbol{CX} = 0$. 由于 $\boldsymbol{1}\boldsymbol{D}_k\boldsymbol{Y}$ 始终为正, 所以原始线性规划等价于

$$\begin{aligned} \min\ & z = \boldsymbol{C}\boldsymbol{D}_k\boldsymbol{Y} \\ \text{s.t.}\ & \boldsymbol{A}\boldsymbol{D}_k\boldsymbol{Y} = 0 \\ & \boldsymbol{1}\boldsymbol{Y} = 1 \\ & \boldsymbol{Y} \geqslant 0 \end{aligned}$$

变换后的问题与原始问题有着相同的格式, 因此我们可以从单纯形的中心 $\boldsymbol{Y} = (\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n})$ 开始重复迭代过程. 在每一次迭代之后, 我们根据 $\boldsymbol{Y}$ 的值计算原始 $\boldsymbol{X}$ 变量的值.

下面给出如何确定变换之后的问题的一个新的可行解. 在任意的第 k 次迭代处, 问题的形式是:

$$\begin{aligned} \min\ & z = \boldsymbol{C}\boldsymbol{D}_k\boldsymbol{Y} \\ \text{s.t.}\ & \boldsymbol{A}\boldsymbol{D}_k\boldsymbol{Y} = 0 \\ & \boldsymbol{Y}\text{位于 } \alpha r\text{−球内} \end{aligned}$$

因为 αr−球是约束 $\boldsymbol{1}\boldsymbol{X} = 1$ 和 $\boldsymbol{X} \geqslant 0$ 所确定空间的子集, 所以这两个约束可以省略. 因此, 上面问题的最优解一定在梯度 $\boldsymbol{c}_p$(最小化) 的映射方向上, 并且可以写成

$$\boldsymbol{Y}_{\text{new}} = \boldsymbol{Y}_0 - \alpha r \frac{\boldsymbol{c}_p}{||\boldsymbol{c}_p||}$$

其中 $\boldsymbol{Y}_0 = (\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n})^{\mathrm{T}}$, $\boldsymbol{c}_p$ 是映射梯度, 可以表示为

$$\boldsymbol{c}_p = [\boldsymbol{I} - \boldsymbol{P}^{\mathrm{T}}(\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}})^{-1}\boldsymbol{P}](\boldsymbol{C}\boldsymbol{D}_k)^{\mathrm{T}}$$

其中

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{A}\boldsymbol{D}_k \\ \boldsymbol{1} \end{pmatrix}$$

α 的选择是提高算法效率关键的一环. 一般情况下, 我们希望选择 α 尽可能地大, 以便算法能够更快地跳到最优解. 然而如果我们选取的 α 太大, 那么就会很快到达单纯形的边界, 甚至跳出可行区域. 到目前为止, 仍然没有一个统一的方法来解决如何选取 α, Karmarkar 建议使用 $\alpha = \frac{n-1}{3n}$.

Karmarkar 算法的基本步骤是

第 0 步 从可行解 $X_0 = (\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n})$ 开始, 计算 $r = \frac{1}{\sqrt{n(n-1)}}$ 和 $\alpha = \frac{(n-1)}{3n}$.

一般的第 k 步 定义

$$\boldsymbol{D}_k = \mathrm{diag}\{x_{k1}, \cdots, x_{kn}\}, \qquad \boldsymbol{P} = \begin{pmatrix} \boldsymbol{A}\boldsymbol{D}_k \\ \boldsymbol{1} \end{pmatrix}$$

计算

$$\boldsymbol{Y}_{\text{new}} = \left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right)^{\mathrm{T}} - \alpha r \frac{\boldsymbol{c}_p}{||\boldsymbol{c}_p||}$$

$$\boldsymbol{X}_{k+1} = \frac{\boldsymbol{D}_k \boldsymbol{Y}_{\text{new}}}{\boldsymbol{1}\boldsymbol{D}_k \boldsymbol{Y}_{\text{new}}}$$

其中

$$\boldsymbol{c}_p = [\boldsymbol{I} - \boldsymbol{P}^{\mathrm{T}}(\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}})^{-1}\boldsymbol{P}](\boldsymbol{c}\boldsymbol{D}_k)^{\mathrm{T}}$$

例 20.3-2

$$\begin{aligned} \min\ z &= 2x_1 + 2x_2 - 3x_3 \\ \text{s.t.}\ -x_1 - 2x_2 + 3x_3 &= 0 \\ x_1 + x_2 + x_3 &= 1 \\ x_1, x_2, x_3 &\geqslant 0 \end{aligned}$$

这个问题满足内点算法的两个基本条件, 即

$$\boldsymbol{X} = (x_1, x_2, x_3)^{\mathrm{T}} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)^{\mathrm{T}}$$

满足两个约束, 最优解

$$\boldsymbol{X}^* = (x_1^*, x_2^*, x_3^*)^{\mathrm{T}} = (0, 0.6, 0.4)^{\mathrm{T}}$$

得到的最优值 $z = 0$.

迭代 0

$$\boldsymbol{c} = (2, 2, -3), \quad \boldsymbol{A} = (-1, -2, 3)$$

$$\boldsymbol{X}_0 = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)^{\mathrm{T}}, \quad z_0 = \frac{1}{3}, \quad r = \frac{1}{\sqrt{6}}, \quad \alpha = \frac{2}{9}$$

$$\boldsymbol{D}_0 = \begin{pmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix}$$

利用映射变换, 可以得到

$$Y_0=\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)^{\mathrm{T}}$$

迭代 1

$$cD_0=(2,2,-3)\begin{pmatrix}\frac{1}{3}&0&0\\0&\frac{1}{3}&0\\0&0&\frac{1}{3}\end{pmatrix}=\left(\frac{2}{3},\frac{2}{3},-1\right)$$

$$AD_0=(-1,-2,3)\begin{pmatrix}\frac{1}{3}&0&0\\0&\frac{1}{3}&0\\0&0&\frac{1}{3}\end{pmatrix}=\left(-\frac{1}{3},-\frac{2}{3},1\right)$$

$$(PP^{\mathrm{T}})^{-1}=\left(\begin{pmatrix}-\frac{1}{3}&-\frac{2}{3}&1\\1&1&1\end{pmatrix}\begin{pmatrix}-\frac{1}{3}&1\\-\frac{2}{3}&1\\1&1\end{pmatrix}\right)^{-1}=\begin{pmatrix}\frac{9}{14}&0\\0&\frac{1}{3}\end{pmatrix}$$

$$\begin{aligned}I-P^{\mathrm{T}}(PP^{\mathrm{T}})^{-1}P&=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}-\begin{pmatrix}-\frac{1}{3}&1\\-\frac{2}{3}&1\\1&1\end{pmatrix}\begin{pmatrix}\frac{9}{14}&0\\0&\frac{1}{3}\end{pmatrix}\begin{pmatrix}-\frac{1}{3}&-\frac{2}{3}&1\\1&1&1\end{pmatrix}\\&=\frac{1}{42}\begin{pmatrix}25&-20&-5\\-20&16&4\\-5&4&1\end{pmatrix}\end{aligned}$$

因此,

$$\begin{aligned}c_p&=\left(I-P^{\mathrm{T}}(PP^{\mathrm{T}})^{-1}P\right)(cD_0)^{\mathrm{T}}\\&=\frac{1}{42}\begin{pmatrix}25&-20&-5\\-20&16&4\\-5&4&1\end{pmatrix}\begin{pmatrix}\frac{2}{3}\\\frac{2}{3}\\-1\end{pmatrix}=\frac{1}{126}\begin{pmatrix}25\\-20\\-5\end{pmatrix}\end{aligned}$$

由此可以得到:

$$\|c_p\|=\sqrt{\frac{25^2+(-20)^2+(-5)^2}{126^2}}=0.257\ 172$$

因此,

$$\begin{aligned}Y_{\text{new}}&=\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)^{\mathrm{T}}-\frac{2}{9}\times\frac{1}{\sqrt{6}}\times\frac{1}{0.257\ 172}\times\frac{1}{126}(25,-20,-5)^{\mathrm{T}}\\&=(0.263\ 340,0.389\ 328,0.347\ 332)^{\mathrm{T}}\end{aligned}$$

接着,

$$\mathit{1}D_0Y_{\text{new}}=\frac{1}{3}(1,1,1)(0.263\ 340,0.389\ 328,0.347\ 332)^{\mathrm{T}}=\frac{1}{3}$$

现在,

$$X_1 = \frac{D_0 Y_{\text{new}}}{\mathbf{1} D_0 Y_{\text{new}}} = \frac{\frac{1}{3} Y_{\text{new}}}{\frac{1}{3}} = Y_{\text{new}} = (0.263\ 340, 0.389\ 328, 0.347\ 332)^{\mathrm{T}}$$

$$z_1 = 0.263\ 34$$

迭代 2

$$cD_1 = (2, 2, -3)\begin{pmatrix} 0.263\ 340 & 0 & 0 \\ 0 & 0.389\ 328 & 0 \\ 0 & 0 & 0.347\ 332 \end{pmatrix}$$
$$= (0.526\ 680, 0.778\ 656, -1.041\ 996)$$

$$AD_1 = (-1, -2, 3)\begin{pmatrix} 0.263\ 340 & 0 & 0 \\ 0 & 0.389\ 328 & 0 \\ 0 & 0 & 0.347\ 332 \end{pmatrix}$$
$$= (-0.263\ 340, -0.778\ 656, 1.041\ 996)$$

$$(PP^{\mathrm{T}})^{-1} = \left(\begin{pmatrix} -0.263\ 340 & -0.778\ 656 & 1.041\ 996 \\ 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} -0.263\ 340 & 1 \\ -0.778\ 656 & 1 \\ 1.041\ 996 & 1 \end{pmatrix}\right)^{-1}$$
$$= \begin{pmatrix} 0.567\ 727 & 0 \\ 0 & 0.333\ 333 \end{pmatrix}$$

$$I - P^{\mathrm{T}}(PP^{\mathrm{T}})^{-1}P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$
$$-\begin{pmatrix} -0.263\ 340 & 1 \\ -0.778\ 656 & 1 \\ 1.041\ 996 & 1 \end{pmatrix}\begin{pmatrix} 0.567\ 727 & 0 \\ 0 & 0.333\ 333 \end{pmatrix}\begin{pmatrix} -0.263\ 340 & -0.778\ 656 & 1.041\ 996 \\ 1 & 1 & 1 \end{pmatrix}$$
$$= \begin{pmatrix} 0.627\ 296 & -0.449\ 746 & -0.177\ 550 \\ -0.449\ 746 & 0.322\ 451 & 0.127\ 295 \\ -0.177\ 550 & 0.127\ 295 & 0.050\ 254 \end{pmatrix}$$

因此,

$$c_p = (I - P^{\mathrm{T}}(PP^{\mathrm{T}})^{-1}P)(cD_1)^{\mathrm{T}}$$
$$= \begin{pmatrix} 0.627\ 296 & -0.449\ 746 & -0.177\ 550 \\ -0.449\ 746 & 0.322\ 451 & 0.127\ 295 \\ -0.177\ 550 & 0.127\ 295 & 0.050\ 254 \end{pmatrix}\begin{pmatrix} 0.526\ 680 \\ 0.778\ 656 \\ -1.041\ 996 \end{pmatrix} = \begin{pmatrix} 0.165\ 193 \\ -0.118\ 435 \\ -0.046\ 757 \end{pmatrix}$$

由此可以得到,

$$\|c_p\| = \sqrt{0.165\ 193^2 + (-0.118\ 435)^2 + (-0.046\ 757)^2} = 0.208\ 571$$

因此,

$$\begin{aligned}\boldsymbol{Y}_{\text{new}} &= \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)^{\mathrm{T}} - \frac{2}{9} \times \frac{1}{\sqrt{6}} \times \frac{1}{0.208\ 571} \times (0.165\ 193, -0.118\ 435, -0.046\ 757)^{\mathrm{T}} \\ &= (0.261\ 479, 0.384\ 849, 0.353\ 671)^{\mathrm{T}}\end{aligned}$$

接着,

$$\boldsymbol{D}_1\boldsymbol{Y}_{\text{new}} = \begin{pmatrix} 0.263\ 340 & 0 & 0 \\ 0 & 0.389\ 328 & 0 \\ 0 & 0 & 0.347\ 332 \end{pmatrix} \begin{pmatrix} 0.261\ 479 \\ 0.384\ 849 \\ 0.353\ 671 \end{pmatrix} = \begin{pmatrix} 0.068\ 858 \\ 0.149\ 832 \\ 0.122\ 841 \end{pmatrix}$$

$$\boldsymbol{1}\boldsymbol{D}_1\boldsymbol{Y}_{\text{new}} = 0.341\ 531$$

现在,

$$\boldsymbol{X}_2 = \frac{\boldsymbol{D}_1\boldsymbol{Y}_{\text{new}}}{\boldsymbol{1}\boldsymbol{D}_1\boldsymbol{Y}_{\text{new}}} = (0.201\ 616, 0.438\ 707, 0.359\ 677)^{\mathrm{T}}$$

$$z_1 = 0.201\ 615$$

重复上面的过程, 所得到的解将越来越接近最优解 $(0, 0.6, 0.4)$. 同时 Karmarkar 也应用了额外的一个步骤将最优解取整到最优极点.

习题 20.3A

1. 使用 TORA 说明: 在例 20.3-1 最后给出的变换了的线性规划问题的解, 确实是原始问题的最优原始对偶解. (提示: 使用 $M = 10$, 并且确保 TORA 的输出精确到小数点之后 5 位.)
2. 将下面的线性规划问题变换成 Karmarkar 求解的格式.

$$\begin{aligned}\max\ & z = y_1 + 2y_2 \\ \text{s.t.}\ & y_1 - y_2 \leqslant 2 \\ & 2y_1 + y_2 \leqslant 4 \\ & y_1,\ y_2 \geqslant 0\end{aligned}$$

3. 将例 20.3-3 再执行一次迭代, 说明得到的解更加接近最优值 $z = 0$.
4. 对于下面的线性规划问题, 执行两次 Karmarkar 算法的迭代:

$$\begin{aligned}\max\ & z = 2x_1 + x_2 \\ \text{s.t.}\ & x_1 + x_2 \leqslant 4 \\ & x_1,\ x_2 \geqslant 0\end{aligned}$$

(提示: 首先将问题转化成 Karmarkar 算法要求的格式.)

5. 对于下面的线性规划问题, 执行两次 Karmarkar 算法的迭代:

$$
\begin{aligned}
\max\ z = {} & 4x_1 + x_3 + x_4 \\
\text{s.t.}\ & -2x_1 + 2x_2 + x_3 - x_4 = 0 \\
& x_1 + x_2 + x_3 + x_4 = 1 \\
& x_1,\ x_2,\ x_3,\ x_4 \geqslant 0
\end{aligned}
$$

(提示: 首先将问题转化成 Karmarkar 算法要求的格式.)

参 考 文 献

Ahuja, R., T. Magnati, and J. Orlin, *Network Flows: Theory, Algorithms, And Applications*, Prentice Hall, Upper Saddle River, NJ, 1993.

Bazaraa, M., J. Jarvis, and H. Sherali, *Linear Programming and Network Flow*, 2nd ed., Wiley, New York, 1990.

Charnes. A. and W. Cooper, "Some Network Characterization for Mathematical Programming and Accounting Applictions to Planning and Control," *The Accounting Review*, Vol. 42, No. 3, pp. 24-52, 1967.

Hooker, J., "Karmarkar's Linear Programming Algorithm," *Interfaces*, Vol. 16, No. 4, pp. 75-90, 1986.

Lasdon, L., *Optimization for Large Systems*, Macmillan, New York, 1970.

第21章　预 测 模 型

本章导读　在决策中, 我们往往要规划未来, 因此描述决策问题的数据就必须要代表未来将会发生的情况. 例如, 在库存控制问题中, 我们根据需求的性质对某个计划期间的生产量做出决策. 再如, 在财务计划中, 我们需要预测现金流随时间的变化情况. 本章介绍 3 种预测技术, 即移动平均、指数平滑和回归, 并给出相应的 Excel 电子表格计算程序.

本章包括 3 个例题、3 个 Excel 应用、9 个节后习题, 以及 1 个放在附录 E 中的案例. AMPL/Excel/Solver/TORA 程序放在 ch21Files 文件夹中.

21.1　移动平均技术

移动平均预测技术假定时间序列是平稳的, 其数据 y_t 对于时间期间 t 可以表示为

$$y_t = b + \varepsilon_t$$

其中

b = 从历史数据中由估计得到的未知常参数

ε_t = 时间 t 期间的随机 (噪声) 项 (均值为零, 方差为常数)

这一技术假设不同期间的数据是相互无关的.

移动平均技术假定, 最新的 n 个观察项对于估计参数 b 是同等重要的. 因此, 在当前期间 t, 若最新的 n 个期间的数据为 $y_{t-n+1}, y_{t-n+2}, \cdots, y_t$, 则期间 $t+1$ 的估计值按下式计算:

$$y^*_{t+1} = \frac{y_{t-n+1} + y_{t-n+2} + \cdots + y_t}{n}$$

对于如何选择移动平均基数 n, 没有严格的规定. 若变量随时间的变化保持为合理的常数, 则建议采用较大的 n; 否则的话, 若数据变化较大, 则 n 的值应该小. 实际中, n 的值通常取 2 到 10 之间.

例 21.1-1

某库存产品过去 24 个月的需求量 (件数) 如表 21.1 所示. 用移动平均技术预测下个月 ($t = 25$) 的需求.

为了选择一个“合理的”期间数 n 的值来计算移动平均值, 首先考察一下已知数据的特点. 表 21.1 的数据表明, 需求量 y_t 有随时间增长的趋势. 一般来说, 这种

趋势意味着, 对于比较长的计划期间, 用移动平均法并不好. 特别是, 在这种情况下不建议用比较大的基数 n 来进行移动平均预测, 因为它会抑制这种数据趋势. 反过来, 假如用较小的 n, 我们就能更好地抓住数据中的这种趋势特性.

表　21.1

月 t	需求量 y_t	月 t	需求量 y_t
1	46	13	54
2	56	14	42
3	54	15	64
4	43	16	60
5	57	17	70
6	56	18	66
7	67	19	57
8	62	20	55
9	50	21	52
10	56	22	62
11	47	23	70
12	56	24	72

如果采用 $n = 3$, 所估计的下个月 $(t = 25)$ 的需求量将等于第 22 到 24 月需求量的平均值, 即

$$y_{25}^* = y_{24}^* = 68 \text{ 件}$$

对于 $t = 25$ 所估计的 68 件的需求也用作 $t = 26$ 的估计, 即

$$x_{26}^* = \frac{70 + 72 + 68}{3} = 70 \text{ 件}$$

当已知 $t = 25$ 的实际需求量时, 应该用它重新估计 23, 24, 25 期间的平均值即 $t = 26$ 时的需求量.

Excel 程序

移动平均法的计算属于 Excel 统计程序包的一部分. 要使用这个程序, 首先要把数据输入到某个单一列的各个单元格, 然后从 Excel 菜单栏上选择 [illegible] ⇒ [illegible] ⇒ 移动平均 来得到一个对话框, 指定存放数据的单元格, 以及存放数据结果的单元格.

图 21.1 显示了 Excel 的移动平均法在求解例 21.1-1(文件 excelEx21.1-1.xls) 问题中的应用. 上部分是对话框, 可用来指定图形输出结果.

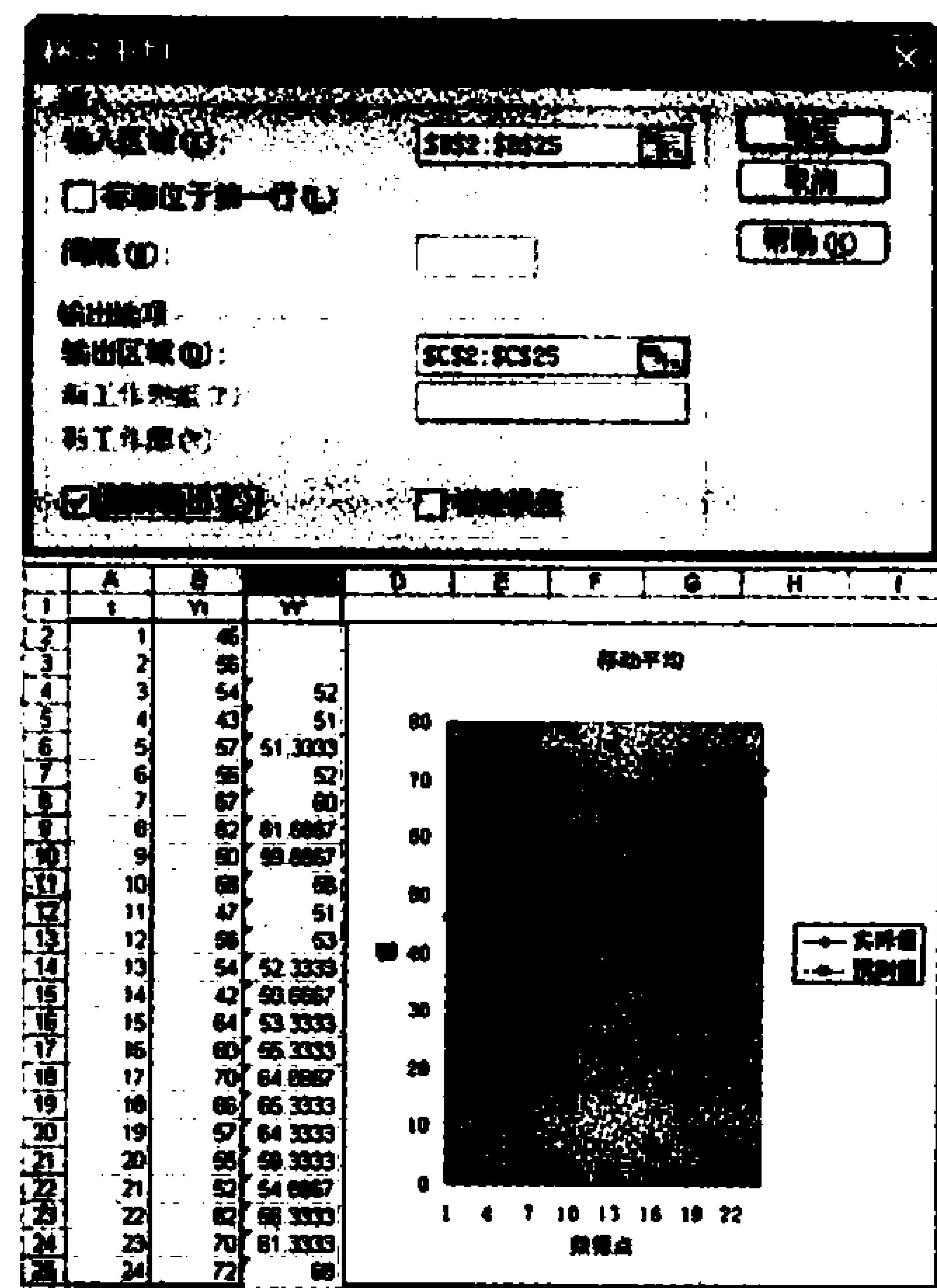


图 21.1 用 Excel 移动平均法程序求解例 21.1-1(文件 excelEx21.1-1.xls)

习题 21.1A

1. 在例 21.1-1 中, 按照 $n = 12$ 估计 $t = 25$ 时的需求量. 较大 n 值对抑制数据的趋势有什么效果?
2. 过去 24 个月空调机的销售量 (台) 如表 21.2 所示. 从移动平均技术实用性的角度对数据进行分析.

表 21.2

月	销售额	月	销售额
1	25	13	40
2	15	14	35
3	30	15	50
4	38	16	60
5	58	17	66
6	62	18	90
7	85	19	105
8	88	20	85
9	60	21	60
10	40	22	55
11	40	23	50
12	38	24	45

3. 表 21.3 给出了 10 年期间游客乘车和乘飞机来到某旅游点的人数. 从移动平均技术实用性的角度对数据进行分析.

表 21.3

年度	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
乘车	1 042	1 182	1 224	1 338	1 455	1 613	1 644	1 699	1 790	1 885
乘飞机	500	522	540	612	715	790	840	900	935	980

4. 表 21.4 是某百货公司的年度销售额 (100 万美元) 数据. 从移动平均技术实用性的角度对数据进行分析.

表 21.4

年度	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
销售额	21.0	23.2	23.2	24.0	24.9	25.6	26.6	27.4	28.5	29.6

5. 某大学在所在州 5 个地点提供校外教学课程. 表 21.5 给出了以往 6 年里的学生人数数据. 每年的数据又按学期分类, 分成秋季 (1), 春季 (2), 夏季 (3). 请你利用这些数据估计下一年的学生人数, 并从移动平均技术实用性的角度对这些数据进行分析.

表 21.5

学期		授课地点				
		1	2	3	4	5
1989						
	1	288	136	48	165	59
	2	247	150	49	168	46
	3	117	69	14	61	15
1990						
	1	227	108	41	108	28
	2	239	106	46	128	43
	3	101	50	15	54	16
1991						
	1	240	126	31	104	46
	2	261	134	19	83	38
	3	138	48	9	56	32
1992						
	1	269	149	17	90	51
	2	301	113	25	54	28
	3	119	50	14	17	6
1993						
	1	226	102	22	16	30
	2	241	110	16	0	24
	3	125	46	7	0	12
1994						
	1	231	88	2	0	1
	2	259	66	3	0	27
	3	102	23	0	0	0

21.2 指数平滑

指数平滑技术采用和移动平均法一样的假定, 即数据过程是常数或变化很小. 但是, 这种技术的设计是为了克服移动平均法的缺点. 具体来说, 指数平滑法对最新的观察数据给出更大的权重, 而移动平均法中对所有的观察数据的权重是一样的.

定义 $\alpha(0<\alpha<1)$ 为**平滑常数**, 并假设过去 t 个期间的时间序列点为 $y_1, y_2, \cdots, y_t$, 则 $t+1$ 期间的估计值 y_{t+1}^* 计算为

$$y_{t+1}^* = \alpha y_t + \alpha(1-\alpha)y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \cdots$$

由于 $y_t, y_{t-1}, y_{t-2}, \cdots$ 的相应系数逐渐减小, 这种新的方法对越近的数据点所给的权重越大.

计算 y_{t+1}^* 的公式可简化如下:

$$\begin{aligned} y_{t+1}^* &= \alpha y_t + (1-\alpha)\{\alpha y_{t-1} + \alpha(1-\alpha)y_{t-2} + \alpha(1-\alpha)^2 y_{t-3} + \cdots\} \\ &= \alpha y_t + (1-\alpha)y_t^* \end{aligned}$$

按照这种方法, y_{t+1}^* 可递归地从 y_t^* 算出, 递归公式开始不用估计 $t=1$ 时的 y_1^*, 让 $t=2$ 的估计等于 $t=1$ 时的实际数据值, 即 $y_2^*=y_1$. 实际上, 开始计算时可以用任何合理的方法, 例如有人建议在时间序列开始时估计 y_0^* 作为某个"合适的"多个期间的平均值.

平滑常数 α 的选取对估计未来的预测非常重要. α 的值大意味着最近的观察数据具有更大的权重. 实际中 α 的值取在 0.01 和 0.30 之间.

例 21.2-1

令 $\alpha=0.1$, 对例 21.1-1 中的数据运用指数平滑技术.

跳过 y_1^* 开始计算, 假定 $y_2^*=y_1=46$ 件. 为了说明递归计算过程, 我们有

$$\begin{aligned} y_3^* &= \alpha y_2 + (1-\alpha)y_2^* = 0.1\times 56 + 0.9\times 46 = 47 \\ y_4^* &= \alpha y_3 + (1-\alpha)y_3^* = 0.1\times 54 + 0.9\times 47 = 47.7 \end{aligned}$$

文件 excelEx21.2-1.xls 提供了 Excel 的指数平滑方法程序用于例 21.1-1 数据的对话框和输出结果. 采用的计算步骤与移动平均法相同 (21.1 节). 注意到 Excel 使用了**阻尼系数** $(=1-\alpha)$, 它是*平滑常数* $(=\alpha)$ 的补值.

根据上述计算方法, $t=25$ 的估计值计算为

$$y_{25}^* = \alpha y_{24} + (1-\alpha)y_{24}^* = 0.1(72) + 0.9(57.63) = 59.07 \text{ 件}$$

这个估计值与移动平均法给出得值 $(=68)$ 件有明显不同. α 的值越大, 对 $t=25$ 的估计就越接近.

习题 21.2A

1. 用 $\alpha = 0.2$ 的指数平滑法分析下列数据.

   (a) 习题 21.1A 的第 2 题.　　(b) 习题 21.1A 的第 3 题.

   (c) 习题 21.1A 的第 4 题.　　(d) 习题 21.1A 的第 5 题.

21.3 回　归

回归分析是确定一个相关变量 (如对某件产品的需求) 与某个独立变量 (如时间) 之间的关系. 独立变量 x 与相关变量 y 的一般的回归公式为

$$y = b_0 + b_1 x + b_2 x^2 + \cdots + b_n x^n + \varepsilon$$

常数 $b_0, b_1, \cdots, b_n$ 为未知参数, 需要从已有的数据中求出来. 随机误差 ε 的平均值为零, 标准差为常数.

最简单的回归模型假定, 相关变量是时间的线性函数, 即

$$y^* = a + bx$$

常数 a 和 b 用**最小二乘准则** (least-squares criterion) 确定, 使得观察值与估计值之差的平方和最小. 已知第 i 个原始数据点 (y_i, x_i), $i = 1, 2, \cdots, n$, 则观察值与估计值之差的平方和定义为

$$S = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

a 和 b 的值可以通过解下列求 S 最小值的必要条件求出:

$$\frac{\partial S}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\frac{\partial S}{\partial b} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) x_i = 0$$

经过代数运算, 得到下列解:

$$b = \frac{\sum_{i=1}^{n} y_i x_i - n\bar{y}\bar{x}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}, \qquad a = \bar{y} - b\bar{x}$$

其中

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

这些公式表明, 我们需要先计算 b, 然后从 b 来计算 a.

a 和 b 的估计对于 y_i 的任何概率分布都成立, 但若 y_i 服从常数标准差的正态分布, 可以计算出预测方法关于在 $x = x^0$(即$y^0 = a + bx^0$) 处的平均值的置信区间:

$$(a+bx^0) \pm t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\sum\limits_{i=1}^{n}(y_i-y_i^*)^2}{n-2}}\sqrt{\frac{1}{n}+\frac{(x^0-\bar{x})^2}{\sum\limits_{i=1}^{n}x_i^2-n\bar{x}^2}}$$

对于相关变量 y 的未来值 (预测值), 我们感兴趣的是求它的**预测区间** (prediction interval)(而不是关于平均值的置信区间). 如我们期望的那样, 一个未来值的预测区间比关于平均值的置信区间要宽. 事实上, 求预测区间的公式除了第二个平方根下的 $\frac{1}{n}$ 项用 $\frac{(n+1)}{n}$ 替换以外, 和求置信区间的公式是一样的.

可以用下列公式计算*相关系数*, 来检查线性估计值 $y^* = a + bx$ 与原始数据的拟合程度:

$$r = \frac{\sum\limits_{i=1}^{n} y_i x_i - n\bar{y}\bar{x}}{\sqrt{\left(\sum\limits_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum\limits_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

其中 $-1 \leqslant r \leqslant 1$.

若 $r = \pm 1$, 则 x 和 y 之间存在最佳的线性拟合. 一般情况下, $|r|$ 的值越接近于 1, 线性拟合越好. 若 $r = 0$, 则 y 和 x 有可能是独立的, 也可能不是独立的因为两个相关的变量可能造成 $r = 0$.

例 21.3-1

用线性回归模型分析例 21.1-1 的数据, 为方便起见, 采用表 21.6 的数据.

表 21.6

月 x_i	需求 y_i	月 x_i	需求 y_i
1	46	13	54
2	56	14	42
3	54	15	64
4	43	16	60
5	57	17	70
6	56	18	66
7	67	19	57
8	62	20	55
9	50	21	52
10	56	22	62
11	47	23	70
12	56	24	72

根据表 21.6 的数据, 我们有

$$\sum_{i=1}^{24} y_i x_i = 17\,842,\quad \sum_{i=1}^{24} x_i = 300,\quad \sum_{i=1}^{24} x_i^2 = 4\,900,\quad \sum_{i=1}^{24} y_i = 1\,374,\quad \sum_{i=1}^{24} y_i^2 = 80\,254$$

因此,

$$\bar{x}=12.5, \bar{y}=57.25$$

$$b=\frac{17\,842-24\times 57.25\times 12.5}{4\,900-24\times 12.5^2}=0.58$$

$$a=57.25-0.58\times 12.5=50$$

因此估计的需求量为

$$y^*=50+0.58x$$

例如, 当 $x=25$ 时, $y^*=50+0.58(25)=64.5$ 件.

相关系数的计算为

$$r=\frac{17\,842-24\times 57.25\times 12.5}{\sqrt{(4\,900-24\times 12.5^2)(80\,254-24\times 57.25^2)}}=0.493$$

r 的值相对小说明 $y^*=50+0.58x$ 可能对于原始数据的拟合得不好. 正常情况下, 好的拟合要求 $0.75\leqslant |r|\leqslant 1$.

假定我们想要计算给定线性回归方程的 95% 置信区间. 首先需要计算关于拟合线偏差的平方和. 表 21.7 给出了相应的信息.

表 21.7

x	y	y^*	$(y-y^*)^2$	x	y	y^*	$(y-y^*)^2$
1	46	50.58	20.98	13	54	57.54	12.53
2	56	51.16	23.43	14	42	58.12	259.85
3	54	51.74	5.11	15	64	58.70	28.09
4	43	54.32	86.86	16	60	59.28	0.52
5	57	52.90	16.81	17	70	59.86	102.82
6	56	53.48	6.35	18	66	60.44	30.91
7	67	54.06	167.44	19	57	61.02	16.16
8	62	54.64	54.17	20	55	61.60	43.56
9	50	55.22	27.25	21	52	62.18	103.63
10	56	55.80	0.04	22	62	62.76	0.58
11	47	56.38	87.98	23	70	63.34	44.36
12	56	56.96	0.92	24	72	63.92	65.29
$\sum_{i=1}^{24}(y_i-y_i^*)^2=1\,205.64$							

从附录 B 的 t 表中可以查到, $t_{0.025,22}=2.074$. 因此, 所需要的置信区间为

$$(50+0.58x^0)\pm 2.074\sqrt{\frac{1\,205.64}{(24-2)}}\sqrt{\frac{1}{24}+\frac{(x^0-12.5)^2}{4\,900-24\times 12.5^2}}$$

上式可简化为

$$(50+0.58x^0)\pm 15.35\sqrt{0.042+\frac{(x^0-12.5)^2}{1\ 150}}$$

为了说明预测区间的计算, 假定我们想建立一个对下个月 ($x^0=25$) 需求估计的预测区间. 在这种情况下, 系数 0.042 必须用 1.042 替换, 所得到的预测区间是 (64.5 ± 16.66), 等于 $(47.84, 81.16)$. 因此我们说, $x=25$ 时的需求数在 47.84 件和 81.16 件之间的概率为 95%.

Excel 程序

回归分析的计算通常比较麻烦. 幸好我们不需要用手算. Excel 提供了自动进行这些计算的工具. 文件 excelEx21.3-1.xls 展示了如何用电子表格程序进行回归分析. 完整的输出报表包括由 Excel 生成的所有必要的信息. 同样地, Excel 中的回归是 工具 菜单下 数据分析 的一个功能块.

习题 21.3A

1. 用线性回归法分析下列数据.
   (a) 习题 21.1A 的第 2 题. (b) 习题 21.1A 的第 3 题.
   (c) 习题 21.1A 的第 4 题. (d) 习题 21.1A 的第 5 题.
2. 在线性回归中, 证明所有数据点的预测值与估计值之间的差的和恒等于零, 即

$$\sum_{i=1}^{n}(y_i-y_i^*)=0$$

参考文献

Brown, B. L., and R. T. O'Connell, *Forecasting and Time Series: An Applied Approach*, Duxbury Press, Belmont, CA, 1993.

Brown, R. G., *Smoothing, Forecasting, and Prediction of Discrete Time Series*, Prentice Hall, Upper Saddle River, NJ, 1972.

Dent, Warren, and Joseph Swanson, "Managerial forecasting by super sophisticated naive models," *Interfaces*, Vol.6, pp. 67-77, 1975. Also commentary in *Interfaces*, Vol.6, No.1, Part 1, pp. 28-31, 1975.

Montgomery, D., and E, Peck, *Introduction to Linear Regression Analysis*, Wiley, New York, 1991.

Willis, R. E., *A Guide to Forecasting for Planners and Managers*, Prentice Hall, Upper Saddle River, NJ, 1987.

第 22 章　随机动态规划

本章导读　本章假定你已经熟悉了第 9 章的确定性动态规划 (DP) 内容. 随机动态规划的要素与确定性情况下的是一样的, 即随机 DP 也要把问题分解成各个阶段, 每个阶段都有相应的状态和决策备选方案. 不同的是, 状态及收益都具有概率性质, 其最优化准则需要根据期望值或者其他某些概率测度来确定. 随机动态规划特别用在随机库存模型的处理以及马尔可夫决策过程中. 这两部分内容分别在第 15 章和第 23 章中介绍. 本章给出的例子说明了动态规划的随机性质.

本章包括 3 个例子、9 个节后习题, 以及 1 个案例.

22.1　一种机会游戏

一种俄式轮盘赌游戏要求旋转一个轮盘, 轮盘外径标刻着 1 到 n 连续的 n 个数字, 一次旋转后轮盘停止在数字 i 的概率是 p_i. 一个游戏参与者花 \$$x$ 可以旋转轮盘最多 m 次. 游戏者的回报是最后一次旋转所产生数字的两倍. 假定这场游戏 (每次最多旋转 m 次) 重玩足够多次的话, 求游戏者的最优策略.

可以按照下面的定义用动态规划模型构造出这个问题.

(1) 用旋转 i 表示阶段 $i, i=1,2,\cdots,m$.

(2) 每个状态下可能的方案包括要么再旋转轮盘一次, 要么停止这次游戏.

(3) 系统在阶段 i 的状态 j 为最后一次旋转所得到的 1 到 n 中的某一个数.

令

$$f_i(j) = \text{已知游戏处于阶段 (旋转)}i\text{ 并且最后一次旋转的结果是 } j \text{ 的最大期望收益}$$

则

$$\begin{pmatrix}\text{已知最后一次旋转的结果为 } j\\ \text{时阶段 } i \text{ 的期望收益}\end{pmatrix} = \begin{cases} 2j, & \text{若游戏结束} \\ \sum_{k=1}^{n} p_k f_{i+1}(k), & \text{若游戏继续}\end{cases}$$

递归公式可写成

$$f_{m+1}(j) = 2j$$

$$f_i(j) = \max\left\{\begin{array}{ll}\text{结束：} & 2j \\ \text{旋转：} & \sum_{k=1}^{n} p_k f_{i+1}(k)\end{array}\right\} \quad i=2,3,\cdots,m$$

$$f_1(0)=\sum_{k=1}^{n} p_k f_2(k)$$

上述递归方程的合理性在于, 当第一次旋转 $(i=1)$ 时, 因为游戏刚刚开始, 系统的状态为 $j=0$, 因此 $f_1(0)=p_1f_2(1)+p_2f_2(2)+\cdots+p_nf_2(n)$. 最后一次旋转 $(i=m)$ 后, 不论第 m 次旋转的结果 j 为何, 这次游戏必须结束, 因此 $f_{m+1}(j)=2j$.

递归计算从 f_{m+1} 开始, 到 $f_1(0)$ 结束, 因此产生 $m+1$ 个计算阶段. 因为 $f_1(0)$ 是从所有 m 次旋转得到的期望收益, 而且知道游戏的成本是 \$$x$, 所以净收益就是 $f_1(0)-x$.

例 22.1-1

假设这种俄式轮盘赌游戏的外径标刻度为 1 到 5 的数字, 停在数字 i 上的概率 p_i 给定为 $p_1=0.3$, $p_2=0.25$, $p_3=0.2$, $p_4=0.15$, $p_5=0.1$. 游戏者付 \$5 最多可旋转 4 次. 求这 4 次旋转每次的最优策略及其期望净收益.

阶段 5

$$f_5(j)=2j,\ j=1,2,3,4,\text{ 或 }5$$

决策: 若 $j=1,2,3,4,5$, 结束, $f_5(j)=2j$.

阶段 4

$$\begin{aligned}f_4(j)&=\max\{2j,\ p_1f_5(1)+p_2f_5(2)+p_3f_5(3)+p_4f_5(4)+p_5f_5(5)\}\\&=\max\{2j,\ 0.3\times2+0.25\times4+0.2\times6+0.15\times8+0.1\times10\}\\&=\max\{2j,\ 5\}=\begin{cases}5, & \text{若 } j=1 \text{ 或 } 2\\ 2j, & \text{若 } j=3,4,5\end{cases}\end{aligned}$$

决策: 若 $j=1$ 或 2, 旋转, $f_4(j)=5$; 若 $j=3,4,5$, 结束, $f_4(j)=2j$.

阶段 3

$$\begin{aligned}f_3(j)&=\max\{2j,\ p_1f_4(1)+p_2f_4(2)+p_3f_4(3)+p_4f_4(4)+p_5f_4(5)\}\\&=\max\{2j,\ 0.3\times5+0.25\times5+0.2\times6+0.15\times8+0.1\times10\}\\&=\max\{2j,\ 6.15\}=\begin{cases}6.15, & \text{若 } j=1,2,3\\ 2j, & \text{若 } j=4 \text{ 或 } 5\end{cases}\end{aligned}$$

决策: 若 $j=1,2,3$, 旋转, $f_3(j)=6.15$; 若 $j=4$ 或 5, 结束, $f_3(j)=2j$.

阶段 2

$$\begin{aligned}f_2(j)&=\max\{2j,\ p_1f_3(1)+p_2f_3(2)+p_3f_3(3)+p_4f_3(4)+p_5f_3(5)\}\\&=\max\{2j,\ 0.3\times6.15+0.25\times6.15+0.2\times6.15+0.15\times8+0.1\times10\}\\&=\max\{2j,\ 6.812\,5\}=\begin{cases}6.812\,5, & \text{若 } j=1,2,3\\ 2j, & \text{若 } j=4 \text{ 或 } 5\end{cases}\end{aligned}$$

决策：若 $j = 1, 2, 3$, 旋转, $f_3(j) = 6.8125$; 若 $j = 4$ 或 5, 结束, $f_3(j) = 2j$.

阶段 1

$$\begin{aligned} f_1(0) &= p_1 f_2(1) + p_2 f_2(2) + p_3 f_2(3) + p_4 f_2(4) + p_5 f_2(5) \\ &= 0.3 \times 6.8125 + 0.25 \times 6.8125 + 0.2 \times 6.8125 + 0.15 \times 8 + 0.1 \times 10 \\ &= 7.31 \end{aligned}$$

决策：旋转, $f_1(0) = \$7.31$.

根据上面的计算, 最优解为

旋转次数	最优策略
1	游戏开始, 旋转
2	若第 1 次旋转产生 1, 2, 3, 继续; 否则, 结束游戏
3	若第 2 次旋转产生 1, 2, 3, 继续; 否则, 结束游戏
4	若第 3 次旋转产生 1 或 2, 继续; 否则, 结束游戏
	期望净收益 = \$7.31 − \$5 = \$2.31

习题 22.1A

1. 在例 22.1-1 中, 假设轮盘标记 1 到 8 的数字, 并可等概率地停在任何一个数字上. 假定每次游戏可有 5 次旋转, 请为该游戏确定一个最优策略.

*2. 我想把二手车卖给出价最高的买家. 经过市场调研发现, 我可以等概率地接受 3 种出价：低价约 \$1 050, 中价约 \$1 900, 高价约 \$2 500. 我把卖车广告张贴出去 3 天. 每天结束时, 我会决定是否接受当天的最佳出价. 我接受出价的最优策略应该是什么?

22.2　投资问题

某个人想在以后的 n 年里要在股票市场投资 10^3C 美元. 投资计划要求每年初买进股票, 并在同年末卖出, 然后把累计的钱数 (全部或部分) 在下一年初再投进去. 投资风险用概率回报来表示. 市场调查表明, 投资回报受到 m 个市场条件的影响, 并且条件 k 以概率 p_k, $k = 1, 2, \cdots, m$ 产生回报 r_k(可以是正的、零或负值). 那么该如何投入这 10^3C 美元来实现在 n 年末达到最高的累计收入呢?

定义

x_i = 第 i 年年初可用的资金额, $i = 1, 2, \cdots, n$ $(x_1 = C)$

y_i = 第 i 年年初投入的资金额 $(y_i \leqslant x_i)$

DP 模型的要素可表述为

(1) 用年度 i 表示阶段 i.

(2) 阶段 i 的备选决策方案由 y_i 给出.

(3) 阶段 i 的状态用 x_i 表示.

令

$f_i(x_i) = i, i+1, \cdots, n$ 年的最大期望资金额, 已知第 i 年年初的 x_i.

对于市场条件 k, 我们有

$$x_{i+1} = (1+r_k)y_i + (x_i - y_i) = x_i + r_k y_i, \quad k = 1, 2, \cdots, m$$

给定市场条件 k 发生的概率 p_k, DP 递归方程可写成

$$f_i(x_i) = \max_{0 \leqslant y_i \leqslant x_i} \left\{ \sum_{k=1}^{m} p_k f_{i+1}(x_i + r_k y_i) \right\}, \quad i = 1, 2, \cdots, n$$

其中 $f_{n+1}(x_{n+1}) = x_{n+1}$, 因为第 n 年以后就没有投资了.

对于年度 n, 我们有

$$f_n(x_n) = \max_{0 \leqslant y_n \leqslant x_n} \left\{ \sum_{k=1}^{m} p_k(x_n + r_k y_n) \right\} = x_n + \max_{0 \leqslant y_n \leqslant x_n} \left\{ \left(\sum_{k=1}^{m} p_k r_k \right) y_n \right\}$$

令

$$\bar{r} = \sum_{k=1}^{m} p_k r_k$$

则有

$$y_n = \begin{cases} 0, & \text{如果 } \bar{r} \leqslant 0 \\ x_n, & \text{如果 } \bar{r} > 0 \end{cases}$$

$$f_n(x_n) = \begin{cases} x_n, & \text{如果 } \bar{r} \leqslant 0 \\ (1+\bar{r})x_n, & \text{如果 } \bar{r} > 0 \end{cases}$$

例 22.2-1

在投资模型中, 假设你想在未来 4 年中投资 \$10 000. 有 50% 的机会能让你的资金翻番, 有 20% 的机会保本, 而另外的 30% 的机会你将损失所有的投资额. 请给出最优的投资策略.

使用模型符号, 我们有

$$C = \$10\,000, \ n = 4, \ m = 3$$

$$p_1 = 0.4, \ p_2 = 0.2, \ p_3 = 0.4$$

$$r_1 = 1, \quad r_2 = 0; \quad r_3 = -1$$

阶段 4

$$\bar{r} = 0.5 \times 1 + 0.2 \times 0 + 0.3 \times (-1) = 0.2$$

因此,

$$f_4(x_4) = 1.2x_4$$

得到最优解如下表:

状 态	最 优 解	
	$f_4(x_4)$	y_4^*
x_4	$1.2x_4$	x_4

阶段 3

$$\begin{aligned} f_3(x_3) &= \max_{0\leqslant y_3\leqslant x_3} \{p_1 f_4(x_3+r_1y_3)+p_2 f_4(x_3+r_2y_3)+p_3 f_4(x_3+r_3y_3)\} \\ &= \max_{0\leqslant y_3\leqslant x_3} \{0.5\times 1.2(x_3+y_3)+0.2\times 1.2(x_3+0y_3)+0.3\times 1.2[x_3+(-1)y_3]\} \\ &= \max_{0\leqslant y_3\leqslant x_3} \{1.2x_3+0.24y_3\} = 1.44x_3 \end{aligned}$$

因此可得

状 态	最 优 解	
	$f_3(x_3)$	y_3^*
x_3	$1.44x_3$	x_3

阶段 2

$$\begin{aligned} f_2(x_2) &= \max_{0\leqslant y_2\leqslant x_2} \{p_1 f_3(x_2+r_1y_2)+p_2 f_3(x_2+r_2y_2)+p_3 f_3(x_2+r_3y_2)\} \\ &= \max_{0\leqslant y_2\leqslant x_2} \{0.5\times 1.44(x_2+y_2)+0.2\times 1.44(x_2+0y_2)+0.3\times 1.44[x_2+(-1)y_2]\} \\ &= \max_{0\leqslant y_2\leqslant x_2} \{1.44x_2+0.288y_2\} = 1.728x_2 \end{aligned}$$

因此可得

状 态	最 优 解	
	$f_2(x_2)$	y_2^*
x_2	$1.728x_2$	x_2

阶段 1

$$\begin{aligned} f_1(x_1) &= \max_{0\leqslant y_1\leqslant x_1} \{p_1 f_2(x_1+r_1y_1)+p_2 f_2(x_1+r_2y_1)+p_3 f_2(x_1+r_3y_1)\} \\ &= \max_{0\leqslant y_1\leqslant x_1} \{0.5\times 1.728(x_1+y_1)+0.2\times 1.728(x_1+0y_1)+0.3\times 1.728[x_1+(-1)y_1]\} \\ &= \max_{0\leqslant y_1\leqslant x_1} \{1.728x_1+0.345\,6y_1\} = 2.073\,6x_1 \end{aligned}$$

因此可得

状态	最优解	
	$f_1(x_1)$	y_1^*
x_1	$2.073\ 6x_1$	x_1

因此最优投资策略为：因为对于 $i=1$ 到 4, $y_i^*=x_i$, 所以最优解要求在每年年初投入所有的资金. 第 4 年年底的累积资金共有 $2.073\ 6x_1 = 2.073\ 6(\$10\ 000) = \$20\ 376$.

事实上可以通过归纳法证明, 上述问题在阶段 $i, i=1,2,\cdots,n$ 有下列一般性的解.

$$f_i(x_i)=\begin{cases} x_i, & \text{如果 } \bar{r}\leqslant 0 \\ (1+\bar{r})^{n-i+1}, & \text{如果 } \bar{r}>0 \end{cases}$$

$$y_i=\begin{cases} 0, & \text{如果 } \bar{r}\leqslant 0 \\ x_i, & \text{如果 } \bar{r}>0 \end{cases}$$

习题 22.2A

*1. 在例 22.2-1 中, 假定概率 p_k 和回报率 r_k 在这 4 年里的变化如下表数据, 求最优投资策略.

年度	r_1	r_2	r_3	p_1	p_2	p_3
1	2	1	0.5	0.1	0.4	0.5
2	1	0	−1	0.4	0.4	0.2
3	4	−1	−1	0.2	0.4	0.4
4	0.8	0.4	0.2	0.6	0.2	0.2

2. 一个 10 m^3 的车厢可用来存放 3 种货品, 货品 1, 2, 3 每件的体积分别为 2 m^3, 1 m^3, 3 m^3. 对每种货品的概率需求如下：

件数	需求的概率		
	货品 1	货品 2	货品 3
1	0.5	0.3	0.3
2	0.5	0.4	0.2
3	0.0	0.2	0.5
4	0.0	0.1	0.0

这 3 种货品每件的缺货费用分别是 \$8, \$10, \$15, 车厢内对每种货品应该装多少件?

3. HiTec 公司刚刚开始在 4 年期间内生产若干台超级计算机. 这种新型计算机的年度需求量 D 服从下列分布：

$$p(D=1)=0.5,\ p(D=2)=0.3,\ p(D=3)=0.2$$

根据工厂的生产能力, 每年可生产 3 台计算机, 每台成本 500 万美元. 每年的实际产量并不一定等于需求量. 年末销售不出去的计算机会产生 100 万美元的存放和维护费用. 假如推迟交货 1 年公司则会损失 200 万美元. 过了 4 年以后, HiTec 公司将不会接受新的订单, 但会在第 5 年继续生产, 以满足第 4 年末没有完成的需求量. 请为 HiTec 公司制定最优的生产计划安排.

*4. PackRat 户外用品公司在小石城内有 3 家体育中心. 复活节这一天, 人们都要开展自行车户外运动. 该公司共有 8 辆自行车供出租, 这些自行车分布在 3 个体育中心, 目的是获得最大的收入. 对这些自行车的需求和每小时收取顾客的租金在这 3 个地点都不一样, 如下面的分布表所示:

自行车数	需求的概率		
	中心 1	中心 2	中心 3
0	0.1	0.02	0
1	0.2	0.03	0.15
2	0.3	0.10	0.25
3	0.2	0.25	0.30
4	0.1	0.30	0.15
5	0.1	0.15	0.10
6	0	0.05	0.025
7	0	0.05	0.025
8	0	0.05	0
租金/小时 ($)	6	7	5

PackRat 公司应该如何把这 8 辆自行车分配在这 3 个体育中心呢?

22.3 最大化实现某个目标的事件

22.2 节介绍了如何求最优期望收益的最大化问题, 而另一种常用的目标是求达到某种收益水平概率的最大化. 我们用 22.2 节的投资问题来说明这种新的目标的应用.

像 22.2 节一样, 我们定义同样的*阶段* i, *备选方案* y_i 和*状态* x_i. 新的目标是最大化在第 n 年年底实现总钱数 S 的概率. 定义

$f_i(x_i)=$ 实现总钱数 S 的概率, 给定 x_i 为第 i 年年初可用的资金量,

并且对第 $i, i+1, \cdots, n$ 年实施某个最优策略

因此, 动态规划的递归方程为

$$f_n(x_n)=\max_{0\leqslant y_n\leqslant x_n}\left\{\sum_{k=1}^{m}p_k\mathrm{P}\{x_n+r_ky_n\geqslant S\}\right\}$$

$$f_i(x_i)=\max_{0\leqslant y_i\leqslant x_i}\left\{\sum_{k=1}^{m}p_k f_{i+1}(x_i+r_k y_i)\right\},\quad i=1,2,\cdots,n-1$$

这些递归公式根据的是条件概率法则：

$$\mathrm{P}\{A\}=\sum_{k=1}^{m}\mathrm{P}\{A|B_k\}\mathrm{P}\{B_k\}$$

在本问题中，用 $f_{i+1}(x_i+r_k y_i)$ 代替 $\mathrm{P}\{A|B_k\}$.

例 22.3-1

有人想投资 \$2 000. 可能的投资结果包括：以概率 0.3 让投入的资金翻番，或者以概率 0.7 损失掉所有的投入. 每年年底把投资卖掉，然后下一年年初再全部或部分进行再投资. 连续 3 年重复这一过程，目标是最大化在第 3 年年底实现 \$4 000 的概率. 简单起见，假定所有的投资都是 \$1 000 的倍数.

用模型的符号，我们说 $r_1=1$ 的概率为 0.3, $r_2=-1$ 的概率为 0.7.

阶段 3 在阶段 3, 状态 x_3 可以最小为 \$0, 最大为 \$8 000. 若整个投资都赔进去了，就达到了最小值，而当投资在前面两年中每年年末都翻番，则会达到最大值. 因此阶段 3 的递归方程可以写成

$$f_3(x_3)=\max_{y_3=0,1,\cdots,x_3}\{0.3\mathrm{P}\{x_3+y_3\geqslant 4\}+0.7\mathrm{P}\{x_3-y_3\geqslant 4\}\}$$

其中 $x_3=0,1,\cdots,8$.

表 22.1 给出了阶段 3 的详细计算过程. 所有阴影区域是不可行的，因为不满足条件 $y_3\leqslant x_3$. 另外，在计算过程中我们注意到

$$\mathrm{P}\{x_3+y_3\geqslant 4\}=\begin{cases}0, & 如果\ x_3+y_3<4\\ 1, & 如果\ x_3+y_3\geqslant 4\end{cases}$$

$$\mathrm{P}\{x_3-y_3\geqslant 4\}=\begin{cases}0, & 如果\ x_3-y_3<4\\ 1, & 如果\ x_3-y_3\geqslant 4\end{cases}$$

虽然表 22.1 表明了备选最选解对于 $x_3=1,3,4,5,6,7,8$ 存在，但最优解列值仅给出了最小的最优值 y_3. 这里的假设是，投资者要想达到所需要的目的，不需要投入不必要的资金.

阶段 2

$$f_2(x_2)=\max_{y_2=0,1,\cdots,x_2}\{0.3f_3(x_2+y_2)+0.7f_3(x_2-y_2)\}$$

相应的计算结果如表 22.2 所示.

表 22.1

x_3	$0.3P(x_3+y_3\geqslant 4)+0.7P(x_3-y_3\geqslant 4)$									最优值	
	y_3=0	1	2	3	4	5	6	7	8	f_3	y_3
0	0.3 × 0+ 0.7×0=0									0	0
1	0.3×0+ 0.7×0=0	0.3×0+ 0.7×0=0								0	0
2	0.3×0+ 0.7×0=0	0.3×0+ 0.7×0=0	0.3×1+ 0.7×0=0.3							0.3	2
3	0.3×0+ 0.7×0=0	0.3×1+ 0.7×0=0.3	0.3×1+ 0.7×0=0.3	0.3×1+ 0.7×0=0.3						0.3	1
4	0.3×1+ 0.7×1=1	0.3×1+ 0.7×0=0.3	0.3×1+ 0.7×0=0.3	0.3×1+ 0.7×0=0.3	0.3 × 1+ 0.7×0=0.3					1	0
5	0.3×1+ 0.7×1=1	0.3×1+ 0.7×1=1	0.3×1+ 0.7×0=0.3	0.3×1+ 0.7×0=0.3	0.3×1+ 0.7×0=0.3	0.3×1+ 0.7×0=0.3				1	0
6	0.3×1+ 0.7×1=1	0.3×1+ 0.7×1=1	0.3×1+ 0.7×1=1	0.3×1+ 0.7×0=0.3	0.3×1+ 0.7×0=0.3	0.31+ 0.7×0=0.3	0.3×1+ 0.7×0=0.3			1	0
7	0.3×1+ 0.7×1=1	0.3×1+ 0.7×1=1	0.3×1+ 0.7×1=1	0.3×1+ 0.7×1=1	0.3×1+ 0.7×0=0.3	0.3×1+ 0.7×0=0.3	0.3×1+ 0.7×0=0.3	0.3×1+ 0.7×0=0.3		1	0
8	0.3×1+ 0.7×1=1	0.3×1+ 0.7×1=1	0.3×1+ 0.7×1=1	0.3×1+ 0.7×1=1	0.3×1+ 0.7×1=1	0.3×1+ 0.7×0=0.3	0.3×1+ 0.7×0=0.3	0.3×1+ 0.7×0=0.3	0.3×1+ 0.7×0=0.3	1	0

表 22.2

x_2	$0.3f_3(x_2+y_2)+0.7f_3(x_2-y_2)$					最优值	
	y_2=0	1	2	3	4	f_2	y_2
0	0.3×0+ 0.7×0=0					0	0
1	0.3×0+ 0.7×0=0	0.3×0.3+ 0.7×0=0.9				0.09	1
2	0.3×0.3+ 0.7×0.3=0.3	0.3×0.3+ 0.7×0=0.09	0.3×1+ 0.7×0=0.3			0.30	0
3	0.3×0.3+ 0.7×0.3=0.3	0.3×1+ 0.7×0.3=0.51	0.3×1+ 0.7×0=0.3	0.3×1+ 0.7×0=0.3		0.51	1
4	0.3×1+ 0.7×1=1	0.3×1+ 0.7×0.3=0.51	0.3×1+ 0.7×0.3=0.51	0.3×1+ 0.7×0=0.3	0.3×1+ 0.7×0=0.3	1	0

阶段 1

$$f_1(x_1)=\max_{y_1=0,1,2}\{0.3f_2(x_1+y_1)+0.7f_2(x_1-y_1)\}$$

相应的计算结果如表 22.3 所示.

表 22.3

x_1	$0.3f_2(x_1+y_1)+0.7f_2(x_1-y_1)$			最优值	
	$y_1=0$	1	2	f_1	y_1
2	$0.3\times0.3+0.7\times0.3$ $=0.3$	$0.3\times0.51+0.7\times0.09$ $=0.216$	$0.3\times1+0.7\times0=0.3$ $=0.3$	0.3	0

最优策略按下列方式确定：已知初始投资额 $x_1=\$2\,000$, 阶段 1(表 22.3) 得到 $y_1=0$, 说明第 1 年不需要投资. 第 1 年不投资的决策使得投资者在第 2 年年初时仍有 \$2 000. 从阶段 2(表 22.2) 的 $x_2=2$ 得到 $y_2=0$, 说明第 2 年还不用投资. 接下来, 从阶段 3(表 22.1) 的 $x_3=2$, 得到 $y_3=2$, 这说明在第 3 年投入所有的资金. 实现目标 $S=4$ 的相应最大概率为 $f_1(2)=0.3$.

习题 22.3A

1. 在例 22.3-1 中, 阶段 1 表明, 备选方案的最优值 $y_1=0, y_2=2$. 证明 $y_1=2$ (即第 1 年投入所有的资金) 也能导致实现所需目标的同样的概率.
2. 用下列数据求解例题 22.3-1：投资者的目标是最大化第 3 年年底获得至少 \$6 000 的概率. 投资者手上有资金 \$1 000, 每年这笔钱翻番的概率都是 0.6.

*3. 假定你在一个赌场玩赌博游戏, 在那里你下一笔赌注, 然后抛硬币两次. 对于你赌的每 \$1, 假如两次抛硬币的结果都是正面朝上, 赌场付给你 \$3, 否则的话, 你就输掉投注的钱. 假设

你总共就只有 \$1, 给定目标是让 3 次游戏后你有 \$4 的概率达到最大, 给出游戏的策略.

参 考 文 献

Bertsekas, D.,*Dynamic Programming: Deterministic and Stochastic Models*, Prentice Hall, Upper Saddle River, NJ, 1987.

Cooper, L., and M. Cooper, *Introduction to Dynamic Programming*, Pergamon Press, New York, 1981.

Smith, D., *Dynamic Programming: A Practical Introduction*, Ellis Horwood, London, 1991.

第 23 章　马尔可夫决策过程

本章导读　本章运用动态规划方法求解具有有限个状态的随机决策过程. 状态之间的转移概率用一个马尔可夫链来描述. 这个过程的收益结构是一个矩阵, 它表示了从一个状态转移到另一个状态相应的收益 (或费用). 转移矩阵和收益矩阵都依赖于决策者的决策方案. 目标是确定最优策略, 使得有限或无限多个阶段上的期望收益达到最大. 本章的预备知识包括马尔可夫链基本原理 (第 17 章)、随机动态规划 (第 22 章) 和线性规划 (第 2 章).

本章包括 6 个例题和 14 个节后习题.

23.1　马尔可夫决策问题的范围

我们用园艺师问题 (例 17.1-1) 来详细说明马尔可夫决策过程. 这个例子的基本思路也适用于其他重要的应用领域, 包括库存、替换问题、货币流管理以及水库容积的调节问题等.

为了方便起见, 这里再来看看不施化肥和施用化肥情况下相应的状态转移矩阵 $\boldsymbol{P}^1$ 和 $\boldsymbol{P}^2$. 阶段 1, 2, 3 分别对应于良好、一般和较差的土壤条件.

$$\boldsymbol{P}^1=\begin{pmatrix}0.2 & 0.5 & 0.3\\ 0 & 0.5 & 0.5\\ 0 & 0 & 1\end{pmatrix},\qquad \boldsymbol{P}^2=\begin{pmatrix}0.30 & 0.60 & 0.10\\ 0.10 & 0.60 & 0.30\\ 0.05 & 0.40 & 0.55\end{pmatrix}$$

从决策问题的观点来看, 该园艺师让某个收益函数 (或收益结构) 与从一个状态到另一个状态的转移联系起来. 这个收益函数表示一年期间的收入或损失, 依赖于转移状态的情况. 由于园艺师可以选择施用化肥或不用化肥, 因此收入或者损失依照不同的决策而有所不同. 矩阵 $\boldsymbol{R}^1$ 和 $\boldsymbol{R}^2$ 分别表示与矩阵 $\boldsymbol{P}^1$ 和 $\boldsymbol{P}^2$ 相关的收益函数 (单位: 100 美元).

$$\boldsymbol{R}^1=\|r_{ij}^1\|=\begin{matrix} \\ 1\\ 2\\ 3\end{matrix}\begin{matrix}\begin{matrix}1 & 2 & 3\end{matrix}\\ \begin{pmatrix}7 & 6 & 3\\ 0 & 5 & 1\\ 0 & 0 & -1\end{pmatrix}\end{matrix},\qquad \boldsymbol{R}^2=\|r_{ij}^2\|=\begin{matrix} \\ 1\\ 2\\ 3\end{matrix}\begin{matrix}\begin{matrix}1 & 2 & 3\end{matrix}\\ \begin{pmatrix}6 & 5 & -1\\ 7 & 4 & 0\\ 6 & 3 & -2\end{pmatrix}\end{matrix}$$

$\boldsymbol{R}^2$ 的元素 r_{ij}^2 考虑施用化肥的费用. 例如, 若去年的土壤条件为一般 (状态 2), 而

且今年变成较差 (状态 3) 的话, 它的收益就是 $r_{23}^2 = 0$, 而不施化肥情况下 $r_{23}^1 = 1$. 因此, $\boldsymbol{R}$ 矩阵给出考虑了化肥费用的因素后的净收益.

那么园艺师的决策问题是什么呢? 首先, 我们必须要了解, 园艺作业是否要连续多年, 还是无限下去. 这些情形被称作**有限阶段** (finite-stage) 和**无限阶段** (infinite-stage) 的决策问题. 在这两种情况下, 园艺师都要用到化学试验的结果 (系统的状态) 来决定最佳的行动方案 (施化肥或不施化肥), 使获得的期望收益最大.

园艺师也可能想要评估一下, 对于一个给定的系统状态所采取的具体行动方案所产生的期望收益如何. 例如, 一旦土壤条件变得较差 (状态 3) 就施用化肥. 这种情况下的决策过程称为是用**平稳策略** (stationary policy) 所表示的决策过程.

每个平稳策略都有不同的转移矩阵和收益矩阵, 这些矩阵可以从 $\boldsymbol{P}^1$, $\boldsymbol{P}^2$, $\boldsymbol{R}^1$, $\boldsymbol{R}^2$ 构造出来. 例如, 对于只有当土壤条件较差 (状态 3) 时才施用化肥的平稳策略, 所得到的转移矩阵和收益矩阵分别为

$$\boldsymbol{P} = \begin{pmatrix} 0.20 & 0.50 & 0.30 \\ 0.00 & 0.50 & 0.50 \\ 0.05 & 0.40 & 0.55 \end{pmatrix}, \quad \boldsymbol{R} = \begin{pmatrix} 7 & 6 & 3 \\ 0 & 5 & 1 \\ 6 & 3 & -2 \end{pmatrix}$$

这些矩阵和 $\boldsymbol{P}^1$, $\boldsymbol{R}^1$ 相比只有第 3 行不同, 它们是直接从关于施用化肥的矩阵 $\boldsymbol{P}^2$, $\boldsymbol{R}^2$ 的第 3 行中取过来的.

习题 23.1A

1. 在园艺师模型中, 如果当土壤条件一般或较差时需要采用施用化肥的平稳策略, 请给出此时的矩阵 $\boldsymbol{P}$ 和 $\boldsymbol{R}$.

*2. 给出该园艺师模型的所有平稳策略.

23.2 有限阶段的动态规划模型

假设园艺师计划在 N 年后从园艺业“退休”, 我们想要求出每年的最优的行动方案 (施用化肥, 或不用化肥), 使得在 N 年末获得最高的期望收益.

令 $k = 1, 2$ 代表园艺师的两种行动方案, 用矩阵 $\boldsymbol{P}^k$ 和 $\boldsymbol{R}^k$ 表示方案 k 的转移概率和收益函数矩阵. 虽然它们已在 2.3.1 节中给出, 但为了方便, 这里把它们总结如下.

$$\boldsymbol{P}^1 = \|r_{ij}^1\| = \begin{pmatrix} 0.2 & 0.5 & 0.3 \\ 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \boldsymbol{R}^1 = \|r_{ij}^1\| = \begin{pmatrix} 7 & 6 & 3 \\ 0 & 5 & 1 \\ 0 & 0 & -1 \end{pmatrix}$$

$$\boldsymbol{P}^2 = \|r_{ij}^2\| = \begin{pmatrix} 0.30 & 0.60 & 0.10 \\ 0.10 & 0.60 & 0.30 \\ 0.05 & 0.40 & 0.55 \end{pmatrix}, \quad \boldsymbol{R}^2 = \|r_{ij}^2\| = \begin{pmatrix} 6 & 5 & -1 \\ 7 & 4 & 0 \\ 6 & 3 & -2 \end{pmatrix}$$

园艺师问题可表示成下面的有限阶段的动态规划 (DP) 模型. 为了一般化, 定义

m = 每个阶段 (年) 的状态数 (园艺师问题中 =3)

$f_n(i)$ = 阶段 $n, n+1, \cdots, N$ 的最优期望收益, 已知 i 是第 n 年年初时的系统状态 (土壤条件)

有关 f_n 和 f_{n+1} 的后向递归方程为

$$f_n(i) = \max_k \left\{ \sum_{j=1}^{m} p_{ij}^k \left[r_{ij}^k + f_{n+1}(j) \right] \right\}, \quad n = 1, 2, \cdots, N$$

其中对所有的 j, $f_{N+1}(j) = 0$.

上述方程的依据是, 从阶段 n 的状态 i 到阶段 $n+1$ 的状态 j 所得到的累计收入 $r_{ij}^k + f_{n+1}(j)$ 发生的概率为 p_{ij}^k. 令

$$v_i^k = \sum_{j=1}^{m} p_{ij}^k r_{ij}^k$$

DP 后向递归方程可写成

$$f_N(i) = \max_k \{v_i^k\}$$

$$f_n(i) = \max_i \left\{ v_i^k + \sum_{j=1}^{m} p_{ij}^k f_{n+1}(j) \right\}, \quad n = 1, 2, \cdots, N-1$$

为说明 v_i^k 的计算, 我们考虑不施用化肥的情况 $(k = 1)$.

$$v_1^1 = 0.2 \times 7 + 0.5 \times 6 + 0.3 \times 3 = 5.3$$

$$v_2^1 = 0 \times 0 + 0.5 \times 5 + 0.5 \times 1 = 3$$

$$v_3^1 = 0 \times 0 + 0 \times 0 + 1 \times -1 = -1$$

因此, 若土壤条件良好, 则一步转移就得到那一年的收益是 5.3; 若土壤一般, 则得到 3; 如果土壤较差, 得到 -1.

例 23.2-1

在本例中, 我们用矩阵 $\boldsymbol{P}^1$, $\boldsymbol{P}^2$, $\boldsymbol{R}^1$, $\boldsymbol{R}^2$ 提供的数据来求解园艺师问题, 时间长度为 3 年 $(N = 3)$.

因为在计算中要重复用到 v_i^k 的值, 为方便起见, 我们在下表中给出它们的值. 前面提到过, $k = 1$ 表示“不用化肥”, $k = 2$ 代表“施用化肥”.

i	v_i^1	v_i^2
1	5.3	4.7
2	3	3.1
3	−1	0.4

阶段 3

i	v_i^k		最优解	
	$k=1$	$k=2$	$f_3(i)$	k^*
1	5.3	4.7	5.3	1
2	3	3.1	3.1	2
3	−1	0.4	0.4	2

阶段 2

i	$v_i^k + p_{i1}^k f_3(1) + p_{i2}^k f_3(2) + p_{i3}^k f_3(3)$		最优解	
	$k=1$	$k=2$	$f_2(i)$	k^*
1	$5.3+0.2\times5.3+0.5\times3.1$ $+0.3\times0.4=8.03$	$4.7+0.3\times5.3+0.6\times3.1$ $+0.1\times0.4=8.19$	8.19	2
2	$3+0\times5.3+0.5\times3.1$ $+0.5\times0.4=4.75$	$3.1+0.1\times5.3+0.6\times3.1$ $+0.3\times0.4=5.61$	5.61	2
3	$-1+0\times5.3+0\times3.1$ $+1\times0.4=-0.6$	$0.4+0.05\times5.3+0.4\times3.1$ $+0.55\times0.4=2.13$	2.13	2

阶段 1

i	$v_i^k + p_{i1}^k f_2(1) + p_{i2}^k f_2(2) + p_{i3}^k f_2(3)$		最优解	
	$k=1$	$k=2$	$f_1(i)$	k^*
1	$5.3+0.2\times8.19+0.5\times5.61$ $+0.3\times2.13=10.38$	$4.7+0.3\times8.19+0.6\times5.61$ $+0.1\times2.13=10.74$	10.74	2
2	$3+0\times8.19+0.5\times5.61$ $+0.5\times2.13=6.87$	$3.1+0.1\times8.19+0.6\times5.61$ $+0.3\times2.13=7.92$	7.92	2
3	$-1+0\times8.19+0\times5.61$ $+1\times2.13=1.13$	$0.4+0.05\times8.19+0.4\times5.61$ $+0.55\times2.13=4.23$	4.23	2

这个最优解说明, 对于第 1 年和第 2 年, 不论系统状态 (化学试验所显示的土壤条件) 如何, 园艺师都应该施用化肥 ($k^*=2$). 而到了第 3 年, 只有当系统处于状态 2 和状态 3 时 (一般或较差土壤条件), 才应该施用化肥. 这 3 年的总期望收益为: 若第 1 年的系统状态为良好, $f_1(1)=10.74$; 若一般, $f_1(2)=7.92$; 如果土壤较差, 则 $f_1(3)=4.23$.

评注 有限时间阶段的问题可以用两种方式加以推广. 首先, 转移概率和收益函数不一定对所有的年份都一样, 其次, 可以对各个阶段的期望收益乘以一个折扣系数, 使得 $f_1(i)$ 等于所有阶段的期望收益的当前现值.

第 1 种推广要求收益值 r_{ij}^k 和转换概率 p_{ij}^k 都是阶段 n 的函数, 如下面的 DP

递归方程所示：

$$f_N(i)=\max_k\{v_i^{k,N}\}$$

$$f_n(i)=\max_k\left\{v_i^{k,n}+\sum_{j=1}^{m}p_{ij}^{k,n}f_{n+1}(j)\right\},\quad n=1,2,\cdots,N-1$$

其中

$$v_i^{k,n}=\sum_{j=1}^{m}p_{ij}^{k,n}r_{ij}^{k,n}$$

在第 2 种推广中，给定 $\alpha(<1)$ 为每年的折扣系数，使得从现在往后一年的 D 美元现值为 αD 美元，则新的递归方程就变成了

$$f_N(i)=\max_k\{v_i^k\}$$

$$f_n(i)=\max_k\left\{v_i^k+\alpha\sum_{j=1}^{m}p_{ij}^{k}f_{n+1}(j)\right\},\quad n=1,2,\cdots,N-1$$

习题 23.2A

*1. 一家公司每年对其重要产品的生产情况进行检查，决定是成功 (状态 1) 还是不成功 (状态 2). 公司必须决定是否要通过广告宣传促进这些产品的销售. 下列矩阵 $\boldsymbol{P}^1$ 和 $\boldsymbol{P}^2$ 提供了每年做广告和不做广告的转移概率，相应的收益由矩阵 $\boldsymbol{R}^1$ 和 $\boldsymbol{R}^2$ 给出. 求未来 3 年的最优决策.

$$\boldsymbol{P}^1=\begin{pmatrix}0.9 & 0.1\\ 0.6 & 0.4\end{pmatrix},\quad \boldsymbol{R}^1=\begin{pmatrix}2 & -1\\ 1 & -3\end{pmatrix}$$

$$\boldsymbol{P}^2=\begin{pmatrix}0.7 & 0.3\\ 0.2 & 0.8\end{pmatrix},\quad \boldsymbol{R}^2=\begin{pmatrix}4 & 1\\ 2 & -1\end{pmatrix}$$

2. 某公司可通过广播、电视和报纸对产品做广告. 这 3 种媒体的每周广告费分别是 \$200, \$900, \$300. 公司每周的产品销售量情况可分为 (1) 一般, (2) 良好, (3) 优秀. 相应每种媒体的转移概率如下：

广播

	1	2	3
1	0.4	0.5	0.1
2	0.1	0.7	0.2
3	0.1	0.2	0.7

电视

	1	2	3
1	0.7	0.2	0.1
2	0.3	0.6	0.1
3	0.1	0.7	0.2

报纸

	1	2	3
1	0.2	0.5	0.3
2	0	0.7	0.3
3	0	0.2	0.8

对应的每周收益矩阵为

广播

$$\begin{pmatrix}400 & 520 & 600\\ 300 & 400 & 700\\ 200 & 250 & 500\end{pmatrix}$$

电视

$$\begin{pmatrix}1\,000 & 1\,300 & 1\,600\\ 800 & 1\,000 & 1\,700\\ 600 & 700 & 1\,100\end{pmatrix}$$

报纸

$$\begin{pmatrix}400 & 530 & 710\\ 350 & 450 & 800\\ 250 & 400 & 650\end{pmatrix}$$

求未来 3 周的最优广告策略.

*3. **库存问题**. 一家电器商店每月初下冰箱订单, 立即进货. 每次固定订货费为 \$100, 每台冰箱每月的储存费为 \$5, 缺货损失费约为每台冰箱每月 \$150. 每月需求量服从下面的 pdf.

需求量 x	0	1	2
$p(x)$	0.2	0.5	0.3

该商店的策略是, 每个月的最大库存水平不超过 2 台冰箱. 求

(a) 本问题不同决策方案的转移概率.

(b) 作为系统状态和决策方案的函数的每月期望库存费用.

(c) 未来 3 个月的最优订货策略.

4. 重解第 3 题, 假定下个季度需求 pdf 的变化如下表:

x	月		
	1	2	3
0	0.1	0.3	0.2
1	0.4	0.5	0.4
2	0.5	0.2	0.4

23.3 无穷多阶段模型

求解无穷多个阶段的问题有两种方法. 第 1 种方法要求评估这个决策问题的所有可能的平稳策略, 相当于穷举过程, 只能用于平稳策略数相对少的情况下. 第 2 种方法称为**策略迭代法** (policy iteration), 它用迭代方法求最优策略, 因而通常更有效率.

23.3.1 穷举法

假设决策问题有 S 个平稳策略, 设 $\boldsymbol{P}^s$ 和 $\boldsymbol{R}^s, s=1,2,\cdots,S$ 为相应策略的 (一步) 转移矩阵和收益矩阵. 穷举法的步骤如下.

第 1 步 给定状态 $i, i=1,2,\cdots,m$, 计算策略 s 的期望一步 (一周期) 收益 v_i^s.

第 2 步 对应于策略 s, 计算其转移矩阵 $\boldsymbol{P}^s$ 的长期平稳概率 π_i^s. 若这些概率存在, 则可从下列方程算出:

$$\boldsymbol{\pi}^s\boldsymbol{P}^s=\boldsymbol{\pi}^s$$

$$\pi_1^s+\pi_2^s+\cdots+\pi_m^s=1$$

其中 $\boldsymbol{\pi}^s=(\pi_1^s,\pi_2^s,\cdots,\pi_m^s)$.

第 3 步 用以下公式求策略 s 的每步 (周期) 转移的期望收益:

$$\mathrm{E}^s=\sum_{i=1}^{m}\pi_i^s v_i^s$$

第 4 步 求最优策略 s^*, 使得

$$E^{s^*} = \max_{s}\{E^s\}$$

下面用这一方法求解无限计划期的园艺师问题.

例 23.3-1

本园艺师问题总共有 8 个平稳策略, 如下表所示:

平稳策略 s	行动方案
1	完全不施肥
2	不论状态如何都施化肥
3	若处于状态 1 时施肥
4	若处于状态 2 时施肥
5	若处于状态 3 时施肥
6	若处于状态 1 或状态 2 时施肥
7	若处于状态 1 或状态 3 时施肥
8	若处于状态 2 或状态 3 时施肥

对于策略 3 到策略 8 的矩阵 $\boldsymbol{P}^s$ 和 $\boldsymbol{R}^s$, 可以从策略 1 和策略 2 的转移概率和收益中得到:

$$\boldsymbol{P}^1 = \begin{pmatrix} 0.2 & 0.5 & 0.3 \\ 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{R}^1 = \begin{pmatrix} 7 & 6 & 3 \\ 0 & 5 & 1 \\ 0 & 0 & -1 \end{pmatrix}$$

$$\boldsymbol{P}^2 = \begin{pmatrix} 0.3 & 0.6 & 0.1 \\ 0.1 & 0.6 & 0.3 \\ 0.05 & 0.4 & 0.55 \end{pmatrix}, \quad \boldsymbol{R}^2 = \begin{pmatrix} 6 & 5 & -1 \\ 7 & 4 & 0 \\ 6 & 3 & -2 \end{pmatrix}$$

$$\boldsymbol{P}^3 = \begin{pmatrix} 0.3 & 0.6 & 0.1 \\ 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{R}^3 = \begin{pmatrix} 6 & 5 & -1 \\ 0 & 5 & 1 \\ 0 & 0 & -1 \end{pmatrix}$$

$$\boldsymbol{P}^4 = \begin{pmatrix} 0.2 & 0.5 & 3 \\ 0.1 & 0.6 & 0.3 \\ 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{R}^4 = \begin{pmatrix} 7 & 6 & 3 \\ 7 & 4 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

$$\boldsymbol{P}^5 = \begin{pmatrix} 0.2 & 0.5 & 0.3 \\ 0 & 0.5 & 0.5 \\ 0.05 & 0.4 & 0.55 \end{pmatrix}, \quad \boldsymbol{R}^5 = \begin{pmatrix} 7 & 6 & 3 \\ 0 & 5 & 1 \\ 6 & 3 & -2 \end{pmatrix}$$

$$\boldsymbol{P}^6 = \begin{pmatrix} 0.3 & 0.6 & 0.1 \\ 0.1 & 0.6 & 0.3 \\ 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{R}^6 = \begin{pmatrix} 6 & 5 & -1 \\ 7 & 4 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

$$\boldsymbol{P}^7 = \begin{pmatrix} 0.3 & 0.6 & 0.1 \\ 0 & 0.5 & 0.5 \\ 0.05 & 0.4 & 0.55 \end{pmatrix}, \quad \boldsymbol{R}^7 = \begin{pmatrix} 6 & 5 & -1 \\ 0 & 5 & 1 \\ 6 & 3 & -2 \end{pmatrix}$$

$$\boldsymbol{P}^8 = \begin{pmatrix} 0.2 & 0.5 & 0.3 \\ 0.1 & 0.6 & 0.3 \\ 0.05 & 0.4 & 0.55 \end{pmatrix}, \quad \boldsymbol{R}^8 = \begin{pmatrix} 7 & 6 & 3 \\ 7 & 4 & 0 \\ 6 & 3 & -2 \end{pmatrix}$$

因此可以计算出 v_i^s 的值, 如下表所示.

s	v_i^s		
	$i=1$	$i=2$	$i=3$
1	5.3	3.0	−1.0
2	4.7	3.1	0.4
3	4.7	3.0	−1.0
4	5.3	3.1	−1.0
5	5.3	3.0	0.4
6	4.7	3.1	−1.0
7	4.7	3.0	0.4
8	5.3	3.1	0.4

平稳概率的计算用下列方程得出:

$$\boldsymbol{\pi}^s \boldsymbol{P}^s = \boldsymbol{\pi}^s$$

$$\pi_1 + \pi_2 + \cdots + \pi_m = 1$$

作为一个例子, 考虑 $s=2$. 相应的方程组为

$$\begin{aligned} 0.3\pi_1 + 0.1\pi_2 + 0.05\pi_3 &= \pi_1 \\ 0.6\pi_1 + 0.6\pi_2 + 0.4\pi_3 &= \pi_2 \\ 0.1\pi_1 + 0.3\pi_2 + 0.55\pi_3 &= \pi_3 \\ \pi_1 + \pi_2 + \pi_3 &= 1 \end{aligned}$$

(注意到前 3 个方程中有一个是冗余的.) 所得到的解为

$$\pi_1^2 = \frac{6}{59}, \quad \pi_2^2 = \frac{31}{59}, \quad \pi_3^2 = \frac{22}{59}$$

在此情况下, 期望年度收益为

$$\begin{aligned} \mathrm{E}^2 &= \pi_1^2 v_1^2 + \pi_2^2 v_2^2 + \pi_3^2 v_3^2 \\ &= \left(\frac{6}{59}\right) \times 4.7 + \left(\frac{31}{59}\right) \times 3.1 + \left(\frac{22}{59}\right) \times 0.4 = 2.256 \end{aligned}$$

下表总结了所有平稳策略的 π^s 和 E^s. (虽然不会影响到计算, 但应注意到策略 1, 3, 4, 6 的每一个都有一个吸收状态: 状态 3. 这也是对于所有这些策略都有 $\pi_1 = \pi_2 = 0$ 和 $\pi_3 = 1$ 的原因.)

s	π_1^s	π_2^s	π_3^s	E^s
1	0	0	1	−1
2	$\frac{6}{59}$	$\frac{31}{59}$	$\frac{22}{59}$	**2.256**
3	0	0	1	0.4
4	0	0	1	−1
5	$\frac{5}{154}$	$\frac{69}{154}$	$\frac{80}{154}$	1.724
6	0	0	1	−1
7	$\frac{5}{137}$	$\frac{62}{137}$	$\frac{70}{137}$	1.734
8	$\frac{2}{135}$	$\frac{69}{135}$	$\frac{54}{135}$	2.216

策略 2 具有最大的期望年度收益. 最优长期策略要求, 不论系统的状态如何都施用化肥.

习题 23.3A

1. 对无穷多个周期, 用穷举法求解习题 23.2A 的第 2 题.
2. 对无限长的计划期, 用穷举法求解习题 23.2A 的第 2 题.

*3. 假定计划期无限长, 用穷举法求解习题 23.2A 的第 3 题.

23.3.2 不带折扣的策略迭代方法

为了充分了解穷举方法的难度, 我们假定, 园艺师有 4 种行动方案, 而不是两种: (1) 不施肥, (2) 每季施肥一次, (3) 施肥两次, (4) 施肥 3 次. 在这种情况下, 园艺师总共有 $4^3 = 256$ 个平稳策略. 通过把行动方案数从 2 个增加到 4 个, 平稳策略数目从 8 个指数“猛增”到 256 个. 这不仅难以直接枚举所有的策略, 而且计算量也会大得不可接受. 这就是为什么我们对建立策略迭代方法感兴趣的原因.

在 23.2 节中, 我们已经说明, 对于每个具体的策略, 在阶段 n 的期望总收益可以表示为递归方程:

$$f_n(i) = v_i + \sum_{j=1}^{m} p_{ij} f_{n+1}(j), \quad j = 1, 2, \cdots, m$$

这一递归方程就是建立策略迭代方法的基础, 但是必须对现在的形式做一点修改, 以便我们能够看出这一过程的渐进行为. 定义 η 为剩下来待考虑的阶段数, 这与方程中的 n 是不一样的, 它定义阶段 n. 因此递归方程可写成

$$f_\eta(i) = v_i + \sum_{j=1}^{m} p_{ij} f_{\eta-1}(j), \quad j = 1, 2, \cdots, m$$

注意, f_η 是假定 η 为剩下来要考虑的阶段数时的累计期望收益. 按照这一新的定义, 这一过程的渐进行为可以通过令 $\eta \to \infty$ 来研究.

设

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_m)$$

为转移矩阵 $\boldsymbol{P} = \|p_{ij}\|$ 的平稳状态概率向量, 且

$$\mathrm{E} = \pi_1 v_1 + \pi_2 v_2 + \cdots + \pi_m v_m$$

为 23.3.1 节所计算的每个阶段的期望收益, 则对充分大的 η, 有

$$f_\eta(i) = \eta\mathrm{E} + f(i)$$

其中 $f(i)$ 为常数项, 表示给定状态 i 时 $f_\eta(i)$ 的渐近截距.

由于 $f_\eta(i)$ 为给定状态 i 时 η 个剩余阶段的累计最优收益, 并且 E 是每个阶段的期望收益, 所以我们直观地可以看出来为什么 $f_\eta(i)$ 等于 $\eta\mathrm{E}$ 加上一个说明具体状态 i 的修正因子 $f(i)$. 这一结果假定了 $\eta \to \infty$.

利用这一信息, 现在把递归方程写成

$$\eta\mathrm{E} + f(i) = v_i + \sum_{j=1}^{m} p_{ij}\{(\eta - 1)\mathrm{E} + f(j)\}, \quad i = 1, 2, \cdots, m$$

简化这个方程, 我们得到

$$\mathrm{E} + f(i) - \sum_{j=1}^{m} p_{ij} f(j) = v_i, \quad i = 1, 2, \cdots, m$$

这样, 我们有 $m+1$ 个未知变量 $f(1), f(2), \cdots, f(m)$ 和 E 的 m 个方程.

如 23.3.1 节一样, 我们的目标是求产生 E 的最大值的最优策略. 因为这 m 个方程带有 $m+1$ 个变量, E 的最优值不能一步求出, 而是需要采用一种两步迭代的方法, 即从任意一个策略开始, 求出一个新的策略, 以得到一个更好的 E 值. 这一迭代过程当连续两个策略完全一样时停止.

(1) **求值步骤**. 选择任意策略 s, 利用相应的矩阵 $\boldsymbol{P}^s$ 和 $\boldsymbol{R}^s$, 并随意假定 $f^s(m) = 0$. 求解方程

$$\mathrm{E}^s + f^s(i) - \sum_{j=1}^{m} p_{ij}^s f^s(j) = v_i^s, \quad i = 1, 2, \cdots, m$$

求出未知的 $\mathrm{E}^s, f^s(1), \cdots, f^s(m-1)$. 转到策略改进步骤.

(2) **策略改进步骤**. 对于每个状态 i, 求可能方案 k, 使得

$$\max_{k}\left\{v_i^k+\sum_{j=1}^{m}p_{ij}^k f^s(j)\right\},\quad i=1,2,\cdots,m$$

$f^s(j)$ 的值, $j=1,2,\cdots,m$ 是在求值步骤求出来的, 所得到的状态 1, 2, $\cdots$, m 的最优决策就构成了新的策略 t. 如果 s 与 t 一致, 则 t 就是最优的, 否则, 赋值 $s=t$, 返回到求值步骤.

例 23.3-2

用策略迭代法求解园艺师问题.

让我们随便从不用施肥的策略开始, 相应的矩阵为

$$\boldsymbol{P}=\begin{pmatrix}0.2&0.5&0.3\\0&0.5&0.5\\0&0&1\end{pmatrix},\quad \boldsymbol{R}=\begin{pmatrix}7&6&3\\0&5&1\\0&0&-1\end{pmatrix}$$

求值迭代步骤的方程为

$$\begin{aligned}
&\mathrm{E}+f(1)-0.2f(1)-0.5f(2)-0.3f(3)=5.3\\
&\mathrm{E}+f(2)\qquad\qquad\quad-0.5f(2)-0.5f(3)=3\\
&\mathrm{E}+f(3)\qquad\qquad\qquad\qquad\qquad\quad-f(3)=-1
\end{aligned}$$

若任意地令 $f(3)=0$, 上述方程得到解

$$\mathrm{E}=-1,\quad f(1)=12.88,\quad f(2)=8,\quad f(3)=0$$

接下来, 使用策略改进步骤. 相应的计算如下表所示.

i	$v_i^k+p_{i1}^kf(1)+p_{i2}^kf(2)+p_{i3}^kf(3)$		最优解	
	$k=1$	$k=2$	$f(i)$	k^*
1	$5.3+0.2\times12.88+0.5\times8$ $+0.3\times0=11.876$	$4.7+0.3\times12.88+0.6\times8$ $+0.1\times0=13.36$	13.36	2
2	$3+0\times12.88+0.5\times8$ $+0.5\times0=7$	$3.1+0.1\times12.88+0.6\times8$ $+0.3\times0=9.19$	9.19	2
3	$-1+0\times12.88+0\times8$ $+1\times0=-1$	$0.4+0.05\times12.88+0.4\times8$ $+0.55\times0=4.24$	4.24	2

这个新的策略要求不论状态如何都施肥. 由于这个新的策略和前一个不同, 因此再次进入求值步骤. 相应新策略的的矩阵是

$$\boldsymbol{P}=\begin{pmatrix}0.3&0.6&0.1\\0.1&0.6&0.3\\0.05&0.4&0.55\end{pmatrix},\quad \boldsymbol{R}=\begin{pmatrix}6&5&-1\\7&4&0\\6&3&-2\end{pmatrix}$$

上述矩阵产生下列方程:

$$
\begin{aligned}
E + f(1) - 0.3f(1) - 0.6f(2) - 0.1f(3) &= 4.7 \\
E + f(2) - 0.1f(1) - 0.6f(2) - 0.3f(3) &= 3.1 \\
E + f(3) - 0.05f(1) - 0.4f(2) - 0.55f(3) &= 0.4
\end{aligned}
$$

仍令 $f(3) = 0$, 我们得到解

$$E = 2.26,\quad f(1) = 6.75,\quad f(2) = 3.80,\quad f(3) = 0$$

策略改进步骤的计算如下表.

i	$v_i^k + p_{i1}^k f(1) + p_{i2}^k f(2) + p_{i3}^k f(3)$		最优解	
	$k=1$	$k=2$	$f(i)$	k^*
1	$5.3 + 0.2 \times 6.75 + 0.5 \times 3.80 + 0.3 \times 0 = 8.55$	$4.7 + 0.3 \times 6.75 + 0.6 \times 3.80 + 0.1 \times 0 = 9.01$	9.01	2
2	$3 + 0 \times 6.75 + 0.5 \times 3.80 + 0.5 \times 0 = 4.90$	$3.1 + 0.1 \times 6.75 + 0.6 \times 3.80 + 0.3 \times 0 = 6.06$	6.06	2
3	$-1 + 0 \times 6.75 + 0 \times 3.80 + 1 \times 0 = -1$	$0.4 + 0.05 \times 6.75 + 0.4 \times 3.80 + 0.55 \times 0 = 2.26$	2.26	2

这个无论什么状态都施肥的新策略和前一个策略一样, 因此最后一个策略就是最优的, 迭代过程终止. 这个结论和穷举法得出的一致 (见 23.3.1 节). 但请注意到, 策略迭代方法能迅速收敛到最优策略, 这也是该新方法的一个特点.

习题 23.3B

1. 在习题 23.2A 的第 1 题中, 假设计划期无限长. 用策略迭代方法求解该问题, 并将计算结果与习题 23.3A 的第 1 题相比较.
2. 在习题 23.2A 的第 2 题中, 假设计划期无限长, 用策略迭代方法求解该问题, 并将计算结果与习题 23.3A 的第 2 题相比较.
3. 在习题 23.2A 的第 3 题中, 假设计划期无限长, 用策略迭代方法求解该问题, 并将计算结果与习题 23.3A 的第 3 题相比较.

23.3.3　带有折扣的策略迭代方法

策略迭代算法可以推广到带有折扣的情形. 设折扣系数为 $\alpha(< 1)$, 有限阶段的递归方程可写成 (见 23.2 节):

$$f_\eta(i) = \max_k \left\{ v_i^k + \alpha \sum_{j=1}^m p_{ij}^k f_{\eta-1}(j) \right\}$$

(注意 η 表示后面要执行的阶段数) 可以证明, 当 $\eta \to \infty$ 时 (无限阶段模型), $f_\eta(i) = f(i)$, 这里 $f(i)$ 是系统处于状态 i 并且在无限长时间内运作的期望现值 (折扣后)

的收益. 因此, $f_\eta(i)$ 当 $\eta \to \infty$ 时的长期行为与 η 的取值无关. 这与没有折扣情况下的 $f_\eta(i) = \eta \mathrm{E} + f(i)$ 的情况完全不同. 这一结果是我们意料之中的, 因为在折扣情况下, 未来收益的的效果将会渐近地趋向于零. 实际情况下, 现值 $f(i)$ 当 $\eta \to \infty$ 应趋向于某一常数.

基于上述信息, 策略迭代步骤可以按如下修改.

(1) **求值步骤**. 选择任意策略 s, 利用相应的矩阵 $\boldsymbol{P}^s$ 和 $\boldsymbol{R}^s$, 求解带有 m 个未知变量 $f^s(1), f^s(2), \cdots, f^s(m)$ 的 m 个方程:

$$f^s(i) - \alpha \sum_{j=1}^{m} p_{ij}^s f^s(j) = v_i^s, \quad i = 1, 2, \cdots, m$$

(2) **策略改进步骤**. 对于每个状态 i, 求可能方案 k, 使得

$$\max_k \left\{ v_i^k + \alpha \sum_{j=1}^{m} p_{ij}^k f^s(j) \right\}, \quad i = 1, 2, \cdots, m$$

$f^s(j)$ 的值, $j = 1, 2, \cdots, m$ 由求值步骤得出. 若得到的策略 t 与 s 相同, 则停止, t 是最优的; 否则, 赋值 $s = t$, 返回到求值步骤.

例 23.3-3

我们用折扣系数 $\alpha = 0.6$ 来求解例 23.3-2.

从任一策略开始, 设 $s = \{1, 1, 1\}$, 由相应的矩阵 $\boldsymbol{P}$ 和 $\boldsymbol{R}$(例 23.3-1 中为 $\boldsymbol{P}^1$ 和 $\boldsymbol{R}^1$) 得到方程:

$$\begin{aligned} f(1) & -0.6[\,0.2f(1) + 0.5f(2) + 0.3f(3)] = 5.3 \\ f(2) & -0.6[\qquad\qquad\; 0.5f(2) + 0.5f(3)] = 3 \\ f(3) & -0.6[\qquad\qquad\qquad\qquad\quad\; f(3)] = -1 \end{aligned}$$

解这组方程得到

$$f_1 = 6.61, \quad f_2 = 3.21, \quad f_3 = -2.5$$

策略改进迭代计算如下表所示:

i	$v_i^k + 0.6[p_{i1}^k f(1) + p_{i2}^k f(2) + p_{i3}^k f(3)]$		最优解	
	$k=1$	$k=2$	$f(i)$	k^*
1	$5.3 + 0.6[0.2 \times 6.61 + 0.5 \times 3.21$ $+0.3 \times (-2.5)] = 6.61$	$4.7 + 0.6[0.3 \times 6.61 + 0.6 \times 0.21$ $+0.1 \times (-2.5)] = 6.90$	6.90	2
2	$3 + 0.6[0 \times 6.61 + 0.5 \times 3.21$ $+0.5 \times (-2.5)] = 3.21$	$3.1 + 0.6[0.1 \times 6.61 + 0.6 \times 3.21$ $+0.3 \times (-2.5)] = 4.2$	4.2	2
3	$-1 + 0.6[0 \times 6.61 + 0 \times 3.21$ $+1 \times (-2.5)] = -2.5$	$0.4 + 0.6[0.05 \times 6.61 + 0.4 \times 3.21$ $+0.55 \times (-2.5)] = 0.54$	0.54	2

求值步骤用 P^2 和 R^2(例 23.3-1) 得到下列方程:

$$
\begin{aligned}
f(1) - 0.6[\ 0.3f(1) + 0.6f(2) + 0.1f(3)] &= 4.7 \\
f(2) - 0.6[\ 0.1f(1) + 0.6f(2) + 0.3f(3)] &= 3.1 \\
f(3) - 0.6[\ 0.05f(1) + 0.4f(2) + 0.55f(3)] &= 0.4
\end{aligned}
$$

解这组方程得到

$$f(1) = 8.89, \quad f(2) = 6.62, \quad f(3) = 3.37$$

策略改进迭代计算得到下表:

i	$v_i^k + 0.6[p_{i1}^k f(1) + p_{i2}^k f(2) + p_{i3}^k f(3)]$		最优解	
	$k=1$	$k=2$	$f(i)$	k^*
1	$5.3 + 0.6[0.2 \times 8.89 + 0.5 \times 6.62$ $+0.3 \times 3.37] = 8.96$	$4.7 + 0.6[0.3 \times 8.89 + 0.6 \times 6.62$ $+0.1 \times 3.37] = 8.89$	8.96	1
2	$3 + 0.6[0 \times 8.89 + 0.5 \times 6.62$ $+0.5 \times 3.37] = 6.00$	$3.1 + 0.6[0.1 \times 8.89 + 0.6 \times 6.62$ $+0.3 \times 3.37] = 6.62$	6.62	2
3	$-1 + 0.6[0 \times 8.89 + 0 \times 6.62$ $+1 \times 3.37] = 1.02$	$0.4 + 0.6[0.05 \times 8.89 + 0.4 \times 6.62$ $+0.55 \times 3.37] = 3.37$	3.37	2

由于这个新的策略 $\{1,2,2\}$ 与前面的策略不同, 用 P^3, R^3(例 23.3-1) 再进入求值步骤, 得到下面的方程:

$$
\begin{aligned}
f(1) - 0.6[\ 0.2f(1) + 0.5f(2) + 0.3f(3)] &= 5.3 \\
f(2) - 0.6[\ 0.1f(1) + 0.6f(2) + 0.3f(3)] &= 3.1 \\
f(3) - 0.6[\ 0.05f(1) + 0.4f(2) + 0.55f(3)] &= 0.4
\end{aligned}
$$

解这组方程得到

$$f(1) = 8.97, \quad f(2) = 6.63, \quad f(3) = 3.38$$

策略改进迭代计算得到下表:

i	$v_i^k + 0.6[p_{i1}^k f(1) + p_{i2}^k f(2) + p_{i3}^k f(3)]$		最优解	
	$k=1$	$k=2$	$f(i)$	k^*
1	$5.3 + 0.6[0.2 \times 8.97 + 0.5 \times 6.63$ $+0.3 \times 3.38] = 8.97$	$4.7 + 0.6[0.3 \times 8.97 + 0.6 \times 6.63$ $+0.1 \times 3.38] = 8.90$	8.98	1
2	$3 + 0.6[0 \times 8.97 + 0.5 \times 6.63$ $+0.5 \times 3.38] = 6.00$	$3.1 + 0.6[0.1 \times 8.97 + 0.6 \times 6.63$ $+0.3 \times 3.38] = 6.63$	6.63	2
3	$-1 + 0.6[0 \times 8.97 + 0 \times 6.63$ $+1 \times 3.38] = 1.03$	$0.4 + 0.6[0.05 \times 8.97 + 0.4 \times 6.63$ $+0.55 \times 3.38] = 3.37$	3.37	2

由于新的策略 $\{1,2,2\}$ 和前一个策略完全一样, 因此它是最优的. 注意到折扣导致了不同的最优策略, 它要求当系统状态为良好时不施用肥料.

习题 23.3C

1. 设有折扣系数 $\alpha = 0.9$, 再次求解下列问题.

   (a) 习题 23.3B 的第 1 题. (b) 习题 23.3B 的第 2 题.

   (c) 习题 23.3B 的第 3 题.

23.4 线性规划解

无限状态的马尔可夫决策问题, 无论是否带有折扣情况, 都可以建立线性规划模型并求解. 首先考虑没有折扣的情况.

23.3.1 节说明了, 不带折扣的无限个状态的马尔可夫问题可化为求最优策略 s^*, 对应于求

$$\max_{s\in S}\left\{\sum_{j=1}^{m}\pi_i^s v_i^s \mid \boldsymbol{\pi}^s\boldsymbol{P}^s=\boldsymbol{\pi}^s,\ \pi_1^s+\pi_2^s+\cdots+\pi_m^s=1,\ \pi_i^s\geqslant 0,\ i=1,2,\cdots,m\right\}$$

集合 S 为问题的所有可能的策略集合. 该问题的约束保证了, π_i^s, $i=1,2,\cdots,m$ 代表马尔可夫链 $\boldsymbol{P}^s$ 的平稳状态概率.

在 23.3.1 节中, 该问题是用穷举法求解的. 具体来说, 每个策略 s 都由固定的一组行动确定 (如例 23.3-1 的园艺师问题). 对于同样的这个问题, 我们可以将它表示为一个线性规划问题, 但是我们需要对问题的变量作一些修改, 使得当系统处于状态 i 时, 最优解能够自动地确定最优的行动 k. 所有的最优行动就能够确定最优策略 s^*.

令

$$q_i^k = \text{已知系统处于状态 } i \text{ 时, 选择行动方案 } k \text{ 的条件概率}$$

因此该问题可表示为

$$\begin{aligned}
\max\quad & \mathrm{E}=\sum_{i=1}^{m}\pi_i\left(\sum_{k=1}^{K}q_i^k v_i^k\right)\\
\text{s.t.}\quad & \pi_j=\sum_{i=1}^{m}\pi_i p_{ij},\quad j=1,2,\cdots,m\\
& \pi_1+\pi_2+\cdots+\pi_m=1\\
& q_i^1+q_i^2+\cdots+q_i^K=1,\quad i=1,2,\cdots,m\\
& \pi_i\geqslant 0,\quad q_i^k\geqslant 0,\quad \text{对于所有的 } i,k
\end{aligned}$$

注意到, p_{ij} 是所选策略的函数, 因此也是策略行动方案 k 的函数.

可通过对 q_i^k 作适当的替换, 可以将这个问题转化成一个线性规划. 请注意这一线性规划与 23.3.1 节中原来的模型是等价的, 仅当 $q_i^k=1$ 对于每个 i 恰好只有

一个k 成立, 这样会把求和 $\sum_{k=1}^{K} q_i^k v_i^k$ 化简为 $v_i^{k^*}$, 其中 k^* 表示所选择的最优行动方案. 我们要建立的线性规划自动地表达了这个条件.

定义

$$w_{ik} = \pi_i q_i^k, \quad \text{对于所有的 } i \text{ 和 } k$$

按照定义, w_{ik} 表示状态 i 进行决策 k 的联合概率. 根据概率论,

$$\pi_i = \sum_{k=1}^{K} w_{ik}$$

因此,

$$q_i^k = \frac{w_{ik}}{\sum_{k=1}^{K} w_{ik}}$$

因此我们得到, 限制条件 $\sum_{i=1}^{m} \pi_i = 1$ 可以写成

$$\sum_{i=1}^{m}\sum_{k=1}^{K} w_{ik} = 1$$

此外, 约束条件 $\sum_{k=1}^{K} q_i^k = 1$ 按照我们用 w_{ik} 来定义 q_i^k 的方式自动成立 (请证明!). 因此这个问题可写成

$$\begin{aligned}
\max \quad & \mathrm{E} = \sum_{i=1}^{m}\sum_{k=1}^{K} v_i^k w_{ik} \\
\text{s.t.} \quad & \sum_{k=1}^{K} w_{jk} - \sum_{i=1}^{m}\sum_{k=1}^{K} p_{ij}^k w_{ik} = 0, \quad j = 1, 2, \cdots, m \\
& \sum_{i=1}^{m}\sum_{k=1}^{K} w_{ik} = 1 \\
& w_{ij} \geqslant 0, \ i = 1, 2, \cdots, m; \ k = 1, 2, \cdots, K
\end{aligned}$$

所建立的模型是一个关于 w_{ik} 变量的线性规划, 它的最优解自动保证每个 i 和每个 k 都可得到一个 q_i^k. 首先注意到, 这个线性规划模型有 m 个独立的约束方程 (有一个与 $\pi = \pi P$ 相关的方程是冗余的). 因此该问题必有 m 个基本变量. 可以证明, 对每个 i, 至少有一个 k 使得 w_{ik} 严格大于 0. 从这两个结果中我们可以得到,

$$q_i^k = \frac{w_{ik}}{\sum_{k=1}^{K} w_{ik}}$$

只能取一个二进制值 (0 或 1). (事实上前面的结果也说明了, $\pi_i = \sum_{k=1}^{K} w_{ik} = w_{ik^*}$, 这里 k^* 是对应于 $w_{ik} > 0$ 的行动方案).

例 23.4-1

下面是不带折扣的园艺师问题的线性规划模型:

$$\begin{aligned}
\max\ E = & 5.3w_{11} + 4.7w_{12} + 3w_{21} + 3.1w_{22} - w_{31} + 0.4w_{32} \\
\text{s.t.}\ & w_{11} + w_{12} - (0.2w_{11} + 0.3w_{12} + 0.1w_{22} + 0.05w_{32}) = 0 \\
& w_{21} + w_{22} - (0.5w_{11} + 0.6w_{12} + 0.5w_{21} + 0.6w_{22} + 0.4w_{32}) = 0 \\
& w_{31} + w_{32} - (0.3w_{11} + 0.1w_{12} + 0.5w_{21} + 0.3w_{22} + w_{31} + 0.55w_{32}) = 0 \\
& w_{11} + w_{12} + w_{21} + w_{22} + w_{31} + w_{32} = 1 \\
& w_{ik} \geqslant 0,\quad \forall\, i,\ k
\end{aligned}$$

最优解为 $w_{11} = w_{12} = w_{31} = 0$, $w_{12} = 0.101\ 7$, $w_{22} = 0.525\ 4$, $w_{32} = 0.372\ 9$. 这个结果意味着, $q_1^2 = q_2^2 = q_3^2 = 1$. 因此, 这个最优解对于 $i = 1, 2, 3$ 选择方案 $k = 2$. E 的最优值是 $4.7(0.101\ 7) + 3.1(0.525\ 4) + 0.4(0.372\ 9) = 2.256$. 有意思的是, 正的 w_{ik} 的值恰好等于与例 23.3-1 的穷举法得出的最优策略相应的 π_i 的值. 这也就说明了这两种方法之间的直接关系.

现在考虑带有折扣的马尔可夫决策问题. 在 23.3.2 节中, 这个问题由下面的递归方程表示:

$$f(i) = \max_k \left\{ v_i^k + \alpha \sum_{j=1}^{m} p_{ij}^k f(j) \right\},\quad i = 1, 2, \cdots, m$$

这些方程等价于

$$f(i) \geqslant \alpha \sum_{j=1}^{m} p_{ij}^k f(j) + v_i^k,\quad \forall\, i,\ k$$

若 $f(i)$ 对每个 i 达到其最小值. 现在考虑目标函数

$$\min \sum_{i=1}^{m} b_i f(i)$$

其中 $b_i (> 0, \forall i)$ 是一个任意常数. 可以证明, 给定不等式约束下求这一函数的最优解将得到 $f(i)$ 的最小值. 因此, 该问题可写成

$$\begin{aligned}
\min\ & \sum_{i=1}^{m} b_i f(i) \\
\text{s.t.}\ & f(i) - \alpha \sum_{j=1}^{m} p_{ij}^k f(j) \geqslant v_i^k,\quad \forall\, i,\ k \\
& f(i) \text{对所有的 } i \text{ 无符号限制}
\end{aligned}$$

这样, 它的对偶问题就是

$$
\begin{aligned}
\max\ & \sum_{i=1}^{m}\sum_{k=1}^{K} v_i^k w_{ik} \\
\text{s.t.}\ & \sum_{k=1}^{K} w_{jk} - \alpha \sum_{i=1}^{m}\sum_{k=1}^{K} p_{ij}^k w_{ik} = b_j, \quad j = 1, 2, \cdots, m \\
& w_{ik} \geqslant 0, \quad i = 1, 2, \cdots, m; \quad k = 1, 2, \cdots, K
\end{aligned}
$$

例 23.4-2

考虑给定折扣系数 $\alpha = 0.6$ 的园艺师问题. 若令 $b_1 = b_2 = b_3 = 1$, 对偶线性规划问题可以写成

$$
\begin{aligned}
\max\ & 5.3w_{11} + 4.7w_{12} + 3w_{12} + 3.1w_{22} - w_{31} + 0.4w_{32} \\
\text{s.t.}\ & w_{11} + w_{12} - 0.6[0.2w_{11} + 0.3w_{12} + 0.1w_{22} + 0.05w_{32}] = 1 \\
& w_{21} + w_{22} - 0.6[0.5w_{11} + 0.6w_{12} + 0.5w_{21} + 0.6w_{22} + 0.4w_{32}] = 1 \\
& w_{31} + w_{32} - 0.6[0.3w_{11} + 0.1w_{12} + 0.5w_{21} + 0.3w_{22} + w_{31} + 0.55w_{32}] = 1 \\
& w_{ik} \geqslant 0, \quad \forall\ i,\ k
\end{aligned}
$$

最优解为 $w_{12} = w_{21}$, $w_{31} = 0$, $w_{11} = 1.567\ 8$, $w_{22} = 3.352\ 8$, $w_{32} = 2.814\ 5$. 这个解表明最优策略为 $(1, 2, 2)$.

习题 23.4A

1. 将下列问题表示成线性规划.
   (a) 习题 23.3B 的第 1 题.　　(b) 习题 23.3B 的第 2 题.
   (c) 习题 23.3B 的第 3 题.

参 考 文 献

Derman, C., *Finite State Markovian Decision Processes*, Academic Press, New York, 1970.

Howard, R., *Dynamic Programming and Markov Processes*, MIT Press, Cambridge, MA, 1960.

Grimmet, G., and D. Stirzaker, *Probability and Random Processes*, 2nd ed., Oxford University Press, Oxford, England, 1992.

Kallenberg, O., *Foundations of Modern Probability*, Springer-Verlag, New York, 1997.

Stewart, W., *Introduction to the Numerical Solution of Markov Chains*, Princeton University Press, Princeton, NJ, 1995.

第24章　案 例 分 析

本章导读　本章介绍15个OR应用案例. 对于每个案例, 我们首先描述它的应用问题, 然后进行详细的分析, 包括数据的采集, 数学模型的建立, 利用AMPL、Excel或 TORA 对模型进行求解, 最后对得到的结果给出解释. 有些例子中, 可能出现整数线性规划计算行为的不可预知性, 如计算时间过长得不出最优解, 或需要对原来的模型进行修正, 以克服计算上的困难. 下表列出了本章将要讨论的案例. AMPL/Excel/Solver/TORA 程序放在文件夹 ch24Files 下.

	案　　例	应用领域	分析工具	软　　件
1	利用最优机动加油量制定航空燃油使用计划	航空运输	LP, 启发式算法	Excel, AMPL
2	心脏瓣膜的最优生产	生产计划	LP	AMPL
3	澳大利亚旅游委员会关于旅游产品交易会的会面安排	旅游业	指派模型, 启发式算法	Excel, AMPL
4	节省联邦政府的旅费支出	商务旅行	最短路径问题	TORA, Excel
5	泰国海军征兵最优行船路线及人员指派问题	运输问题 路线安排	运输模型, ILP	AMPL
6	Mount Sinai 医院手术室时间安排	医疗卫生	ILP, GP	AMPL
7	PFG 建材玻璃公司的拖车有效荷载优化	分销行业	自由体受力图, ILP	AMPL
8	木材加工中的最优切割问题	木材厂工艺	DP	Excel
9	计算机集成制造 (CIM) 设施的布局规划问题	布局规划	AHP, GP	Excel, AMPL
10	旅店订房预定上限问题	旅店业	决策树	Excel
11	Casey 问题: 新化验的解释与评估	医疗化验	贝叶斯概率, 决策树	Excel
12	莱德杯决赛中高尔夫球手的出场顺序	体育	对策论	Excel, TORA
13	戴尔公司供应链的库存决策	库存控制	库存模型	Excel
14	某制造厂内部运输系统的分析	原材料管理	排队论, 模拟	Excel, TORA
15	Qantas 航空公司电话售票人力计划	航空业	ILP, 排队论	Excel, AMPL, TORA

案例 1　利用最优机动加油量制定航空公司的燃油使用计划①

工具： 启发式算法, LP

应用领域： 航空运输

问题的描述

通常, 一个商业航班路线会形成一个圈, 从航空公司的运行枢纽机场出发, 经停若干城市, 再回到枢纽站. 例如, 一条路线从洛杉矶 (LAX) 出发, 经过坦帕 (TPA)、迈阿密 (MIA)、劳德代尔 (FLL)、纽约 (LGA)、迈阿密 (MIA) 和休斯顿 (IAH), 最后再回到洛杉矶. 飞机可以在飞行路线上的任何一个机场加油, 但由于在各经停站的燃料费用不同, 通过*机动加油量*(tankering)就可能节约大量的燃料成本. 机动加油量指在某些经停机场充分利用低油价, 加注多余的燃油, 利用这些剩余的燃油飞行后续的航段. 而机动加油量的缺点是, 由于增加了飞机的载重量, 使得汽油的燃烧量也相应增加. 因此, 机动加油量只有在节约的燃油费大于超额燃烧费用的前提下才有意义. 这个问题就变成了考虑飞机载重量限制的前提下, 求飞行路线上的最优的机动加油量问题.

求解机动加油量问题的启发式方法

图 24.1 表示途经 n 个城市的一个飞行路线, 城市 1 表示路线的起点和终点. 用于从经停点 j 到 $j+1$ 的航段 j 上的机动加油量可以在 $1,2,\cdots,j-1$ 的任何点上加注. 定义

$$x_{ij} = \text{在经停站 } i<j \text{ 加注用于飞行航段 } j \text{ 的燃油量比例}$$

根据定义, $\sum_{i=1}^{j-1} x_{ij} \leqslant 1$, $x_{ij} \geqslant 0$. 若 $\sum_{i=1}^{j-1} x_{ij} < 1$, 则机动加油量不能提供航段 j 所需的所有耗油量, 差量 $1-\sum_{i=1}^{j-1} x_{ij}$ 必须在经停点 j 加满.

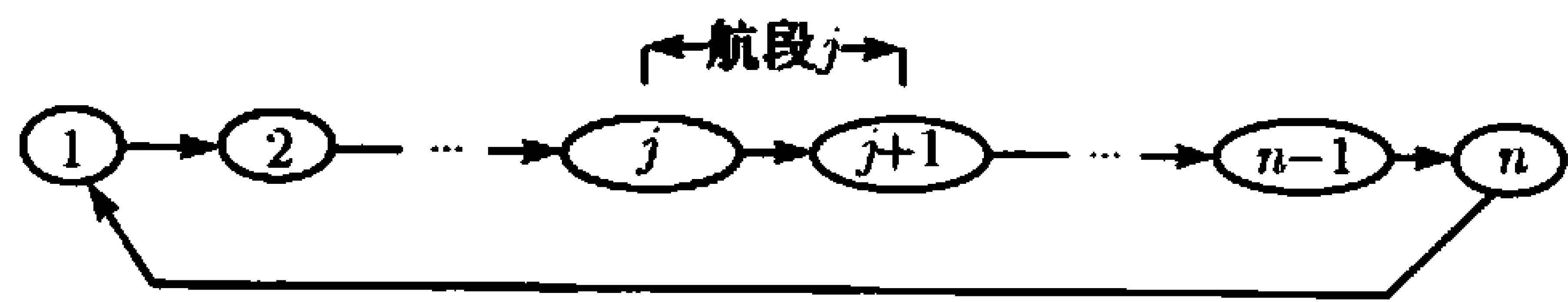


图 24.1　一个经停 n 站的航班路线

启发式方法的基本思路是, 把在经停点 $i=1,2,\cdots,j-1$ 为航段 j 加注的机动加油量的加权油价与在经停点 j 为航段 j 购买燃油的基准价格进行比较. 之所

① **资料来源：** Nash, B. “A Simplified Alternative to Current Airline Fuel Allocation Models,” *Interfaces*, Vol.11, No.1, pp.1-8, 1981.

以要使用机动加油量的加权价格，是因为在一个经停点为后续航段加注燃油会由于载重量增加而导致多余的燃烧量 (多余燃烧惩罚). 令

q = 油箱所带多余燃料每加仑每小时额外燃烧所占的比例数

t_{ij} = 从机动加油的航站 i 到航段 j 的出发航站 $(i < j)$ 的时间小时数

$\hat{p}_j$ = 经停站 j 每加仑燃油的基准价格

加权价格 p_{ij} 可用下列公式计算:

$$p_{ij} = \frac{\hat{p}_i}{(1-q)^{t_{ij}}}, \quad i < j$$

令

$$p_{i^*j} = \min_{i<j}\{p_{ij}\}$$

启发式算法规则建议, 若

$$p_{i^*j} < \hat{p}_j, \quad i^* < j$$

则在航站 i^* 为航段 j 添加机动加油量.

为了说明如何运用这一启发式算法, 图24.2给出了一条航班路线, 从丹佛(DEN) 出发, 途经堪萨斯城 (MCI)、圣路易斯 (STL) 和托皮卡 (FOE), 最后再回到丹佛. 相应的弧上给出了飞行小时数, 节点上标出了每加仑汽油的基准价格.

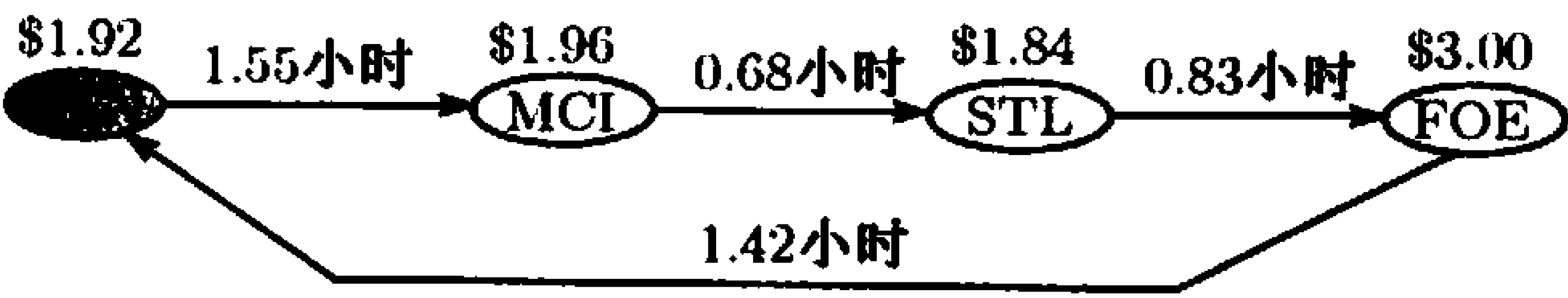


图 24.2　路径 DEN-MCI-STL-FOE-DEN 上的汽油价格及飞行时间

图 24.3 给出了本案例航线的机动加油机会. 相应的数字编码 1, 2, 3, 4 分别代

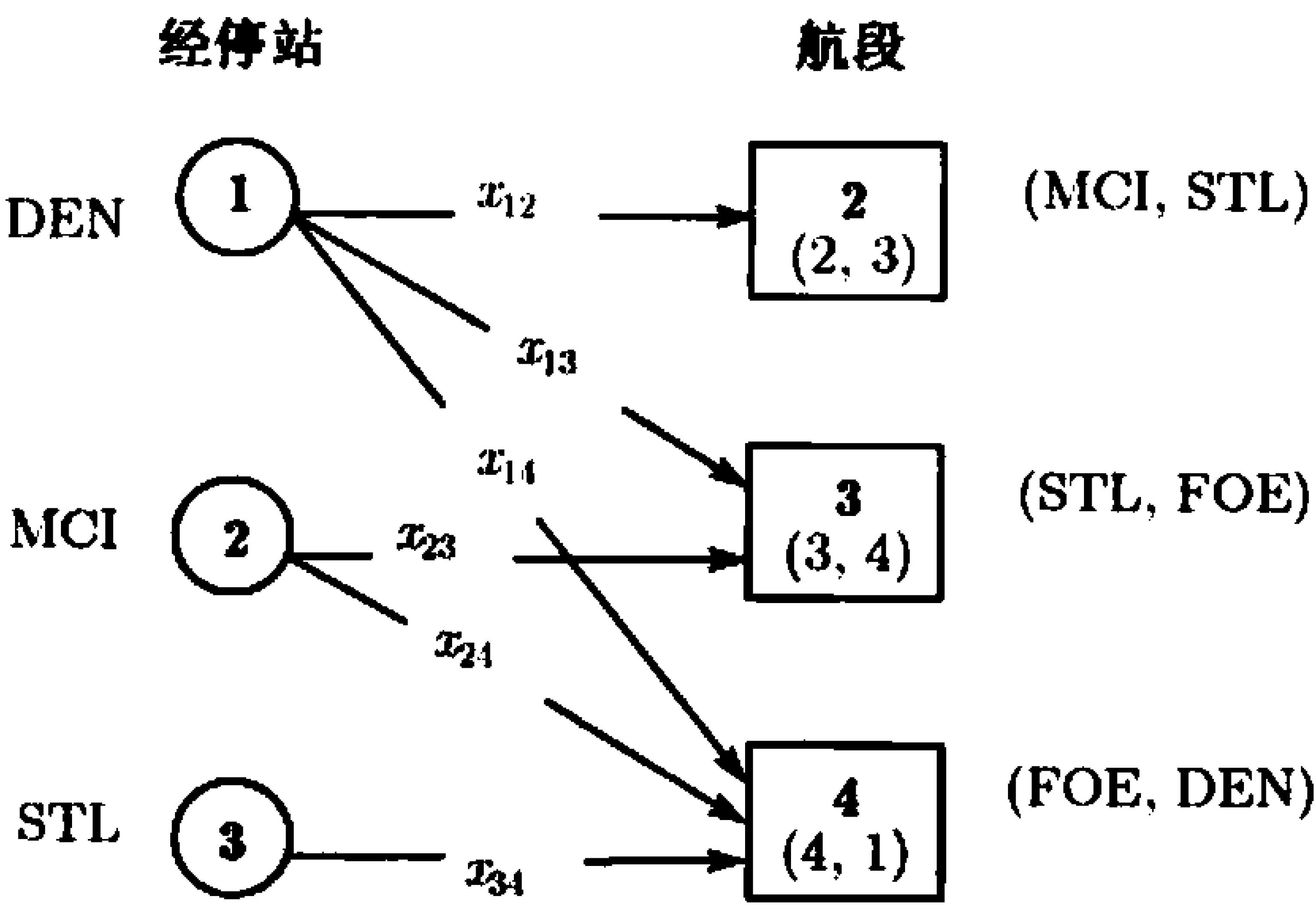


图 24.3　图 24.2 中航班路线的要素的表示

表 DEN, MCI, STL, FOE. 航段 (DEN, MCI), (MCI, STL), (STL, FOE), (FOE, DEN) 也用相应的代码 1, 2, 3, 4 表示. 这个图说明, 经停站 1(DEN) 可为航段 2, 3, 4 进行机动加油, 经停站 2(MCI) 可为航段 3 和 4 加油, 而经停站 3(STL) 只能为航段 4 加油. 因为 DEN 是航班路线圈的起点和终点, 因此航段 1(DEN, MCI) 不存在机动加油问题. 相应的弧上定义了对应的机动加油比 x_{ij}.

可以方便地用电子表格程序 (文件 excelCase1.xls) 来执行启发式算法的计算. 图 24.4 把这个电子表格分成两部分: 输入部分 (3~9 行) 和输出部分 (12~27 行). 输入数据包括 $q(=0.05)$ 值、每个经停站每加仑燃油的基准价格, 以及航段飞行小时数. 输出部分首先计算由一个经停站到一个航段出发航站的机动加油的飞行时间. 例如, 在 DEN(经停站 1) 加注的燃油要飞行 1.55 小时到达 MCI, 航段 2 的开始航站 MCI, 要飞行 $2.23(=1.55+0.68)$ 小时到达第 3 个航段的出发航站 STL, 并且要花 $3.06(=1.55+0.68+0.83)$ 小时到达第 4 个航段的出发航站 FOE. 分段小时数则用来计算加权价格 (19~24 行). 例如, 从 DEN(经停站 1) 为航段 2(MCI, STL) 加油的加权价格计算为

$$p_{12}=\frac{\hat{p}_1}{(1-q)^{t_{12}}}=\frac{1.92}{(1-0.05)^{1.55}}=2.079$$

	A	B	C	D	E
1	Tankering Heuristic				
2	*Input data:*				
3	q	0.05			
4		DEN	MCI	STL	FOE
5	Stopover	1	2	3	4
6	Price per gallon($)	1.92	1.96	1.84	3
7		DEN-MCI	MCI-STL	STL-FOE	FOE-DEN
8	Leg	1	2	3	4
9	Flight hours	1.55	0.68	0.83	1.42
10					
11	*Output:*				
12	Block hours				
13		Leg 2	Leg 3	Leg 4	
14	Stopover 1	1.55	2.23	3.06	
15	Stopover 2		0.68	1.51	
16	Stopover 3			0.83	
17	Stopover 4				
18					
19	Weighted price				
20		Leg 2	Leg 3	Leg 4	
21	stopover 1	2.0788811	2.15267	2.246296	
22	stopover 2		2.02957	2.117841	
23	stopover 3			1.920027	
24	stopover 4				
25	Min weighted price	2.0788811	2.02957	1.920027	
26	Tankering(yes/no)?	no	no	yes	
27	Tankering stopover			STL	

图 24.4 机动加油量启发式算法的电子表格计算程序 (文件 excelCase1.xls)

这些加权价格建议, 只在经停站 3(STL) 为航段 4(FOE, DEN) 进行机动加油, 因为它提供的加权价格 \$1.92 小于在 FOE 的 \$3.00 的基准价格.

所提出的启发式算法作了两条基本假设:

(1) 任何航站的燃料供应量没有限制.

(2) 飞机的承载能力能够接受加注的燃油量.

在这些条件下, 启发式解总是最优的. 在其他情况下, 就必须修改启发式规则, 按下列方法得到一个可行解:

(1) 对候选的机动加油城市按照加权价格的升序进行排队. (一个机动加油经停站的加权价格必须小于目标航段出发航站的基准价格)

(2) 在加权价格最低的经停站分配尽可能多的燃油 (在经停站供应量和飞机承载重量的约束下).

(3) 对排在下一个的经停站重复第 2 步, 直到满足一个航段的需求, 或者直到在不违背经停站供应量和载重量限制的条件下不能进行机动加油为止.

(4) 若某个航段仍不满足燃料需求, 则在该航段的出发航站按照基准价格购买不足的燃油.

为了说明如何应用这一带有容量限制的启发式算法, 考虑表 24.1 中的一组数据, 已知航段 j 的燃料需要量、经停站 j 的机动加油容量 A_j, 以及经停站 j 的燃油供应上限 B_j.

表 24.1 案例 1 数据的例子

航段 $j \equiv (j, j+1)$	航段 j 所需燃油加仑数 f_j	经停站 j 机动装油容量 A_j	经停站 j 供油上限 B_j	经停站 j 每加仑基准价格 $\hat{p}_j$
1 (DEN,MCI)	1 200	860	1 500	\$1.92
2 (MCI,STL)	645	960	900	1.96
3 (STL,FOE)	880	900	1 400	1.84
4 (FOE,DEN)	920	900	1 500	3.00

若不考虑经停点 j 的机动装油容量 A_j 以及油料供应限制 B_j, 图 24.4 的解要求在 STL 为航段 4 加油 $\frac{920}{(1-0.05)^{0.83}} = 961$ 加仑. 这一圈飞行的总燃油费用为 $1\ 200 \times \$1.92 + 645 \times \$1.96 + (880 + 961) \times \$1.84 = \$6\ 955.64$. 另一方面, 如果考虑这些限制的话, 所给出的解就不可行了, 因为在 STL 加油 $880 + 961 = 1\ 841$ 加仑既超过了机动装油容量 ($A_3 = 900$ 加仑), 也超出了供油限制 ($B_3 = 1\ 400$ 加仑). 这意味着在 STL 最多可加油的量为 $900 - 880 = 20$ 加仑, 相当于为航段 4 净加油量 $20(1-0.05)^{0.83} \approx 19$ 加仑, 造成该航段缺油 $920 - 19 = 901$ 加仑. 这个量必须在 FOE 航站购买, 因为飞机在 STL 已经达到了机动装油容量上限 (= 900 加仑), 从而不可能在 DEN 和/或 MCI 进行机动加油. 因此该航线的燃油费用等于 $1\ 200 \times \$1.92 + 645 \times \$1.96 + (20 + 880) \times \$1.84 + 901 \times \$3.00 = \$7\ 927.20$, 总费用比没有限制的解增加了 $7\ 927.20 - 6\ 955.64 = \$971.56 (\approx 14\%)$.

用线性规划求最优解

线性规划模型在下面两类约束下求燃料总支出的最小值: (1) 飞机最大加油量;

(2) 经停航站的供油上限. 决策变量 x_{ij} 表示在经停站 i 为航段 j 所加的燃料油的比例. 令

f_j= 航段 j 所需燃料的净加仑数

F_{ij}= 在经停站 i 为航段 $j(i<j)$ 所加注燃料 (包括额外燃烧量) 的总加仑数

量 F_{ij} 根据 f_j 计算为

$$F_{ij} = \frac{f_j}{(1-q)^{t_{ij}}}, \quad i<j$$

按照前面的定义, q 是每小时每加仑额外燃烧量的比例, t_{ij} 为从加油经停站 i 到航段 j 出发航站的小时数.

在经停站 k 飞机上所加注的燃油量计算为

$$T_k = \sum_{j=k+1}^{n} F_{kj}\left(\sum_{i\leqslant k} x_{ij}\right), \quad k=1,2,\cdots,n-1$$

这个公式表达了在经停站 k 及其以前为所有飞过经停站 k 的后续航段 $k+1, k+2, \cdots, n-1$ 所加的油量 (参见图 24.3).

接下来, 在经停站 k 购买的燃油量计算为

$$Q_k = f_k\left(1-\sum_{i<k} x_{ik}\right) + \sum_{j>k} F_{kj}x_{kj}, \quad k=1,2,\cdots,n$$

在经停站 k 购买的油量包括两部分: (1) 航段 k 还需要添加的额外燃油量, (2) 用于后续航段 $k+1, k+2, \cdots, n$ 所加的燃油量. 注意到在经停站 k 购买用于航段 k 的油量只有 $f_k\left(1-\sum\limits_{i<k} x_{ik}\right)$ 加仑, 而不是 f_k 加仑, 因为有 $f_k\sum\limits_{i<k} x_{ik}$ 的量在前面的经停站 $i<k$ 已经加注过了.

利用前面定义的 A_i 和 B_i, 来表示经停站 i 的机动装油容量和供油上限, 这一线性规划模型表达如下:

$$\begin{aligned}
\min\ & z = \sum_{i=1}^{n} \hat{p}_i Q_i \\
\text{s.t. } & T_i \leqslant A_i, \quad i=1,2,\cdots,n-1 \\
& Q_i \leqslant B_i, \quad i=1,2,\cdots,n \\
& \sum_{i=1}^{j} x_{ij} \leqslant 1, \quad j=2,3,\cdots,n \\
& x_{ij} \geqslant 0, \quad i<j
\end{aligned}$$

这一模型可通过上面给出的用 x_{ij} 表示的公式代入 T_i 和 Q_i 来求解. 注意到第三组约束采用了不等式约束 ($\leqslant$), 而没有用严格等式约束 ($=$), 它表示可以不为航段 j 加油.

表 24.1 中启发式例子的数据可用于说明如何用线性规划模型求解该问题. 第一步用 f_i 和 t_{ij} 计算 F_{ij}, 如表 24.2 所示. 为了方便, 所有的值都被增大到下一个邻近的整数值. 为了说明计算过程, 我们有

$$F_{12}=\frac{f_2}{(1-q)^{t_{12}}}=\frac{645}{(1-0.05)^{1.55}}=698.37\approx 699 \text{ 加仑}$$

表 24.2　用 f_j 和 t_{ij} 计算 F_{ij}

i	F_{ij}		
	$j=2$	$j=3$	$j=4$
1	699	987	1 077
2		912	995
3			961

因此我们得到

$$\begin{aligned}
T_1&=699x_{12}+987x_{13}+1\ 077x_{14}\\
T_2&=912(x_{13}+x_{23})+995(x_{14}+x_{24})\\
T_3&=961(x_{14}+x_{24}+x_{34})\\
Q_1&=1\ 200+699x_{12}+987x_{13}+1\ 077x_{14}\\
Q_2&=645(1-x_{12})+912x_{23}+995x_{24}\\
Q_3&=880(1-x_{13}-x_{23})+961x_{34}\\
Q_4&=920(1-x_{14}-x_{24}-x_{34})
\end{aligned}$$

因此整个线性规划为

$$\begin{aligned}
\min\quad & z=1.92Q_1+1.96Q_2+1.84Q_3+3.00Q_4\\
\text{s.t.}\quad & Q_1\leqslant 1\ 500,\ Q_2\leqslant 900,\ Q_3\leqslant 1\ 400,\ Q_4\leqslant 1\ 500\\
& 699x_{12}+987x_{13}+1\ 077x_{14}\leqslant 860\\
& 912x_{13}+912x_{23}+995x_{14}+995x_{24}\leqslant 960\\
& 961x_{14}+961x_{24}+961x_{34}\leqslant 900\\
& Q_1-699x_{12}-987x_{13}-1\ 077x_{14}=1\ 200\\
& Q_2+645x_{12}-912x_{23}-995x_{24}=645\\
& Q_3+880x_{13}+880x_{23}-961x_{34}=880\\
& Q_4+920x_{14}+920x_{24}+920x_{34}=920\\
& x_{12}\leqslant 1
\end{aligned}$$

$$x_{13} + x_{23} \leqslant 1$$

$$x_{14} + x_{24} + x_{34} \leqslant 1$$

$$x_{ij} \geqslant 0,\ i = 1,2,3;\ j = 2,3,4$$

AMPL 模型

文件 amplCase1.txt 给出了求解机动加油量问题线性规划的 AMPL 模型. 输入内容包括该问题的原始数据 (按照数学模型的定义, 即 $n, q, \hat{p}_i, f_i, A_i, B_i$). 图 24.5 给出了表 24.1 数据的输出结果. 为航段 4(STL-FOE) 所加的油量分别在经停站 1, 2, 3 (DEN, MCI, STL) 购买, 加油量分别为 150(= 1 350 − 1 200), 255(= 900 − 645), 520(= 1 400 − 880) 加仑. 去掉额外燃烧的燃料, 加油购买量减少到 $150(1-0.05)^{(1.55+0.68+0.83)} + 255(1-0.05)^{(0.68+0.83)} + 520(1-0.05)^{0.83} \approx 862$ 加仑. 这个量加上在城市 4(FOE) 所购买的 58 加仑, 得到 862 + 58 = 920 加仑, 这就是航段 4 所需要的燃油量. 如我们所期望的, 最优的 LP 费用比启发式算法得到的要少 \$7 927.20 − \$7 104.44 = \$822.76.

```
Total cost =    7104.44

City   Supply limit   Purchased amount   Min requirement    Cost
 1        1500             1350               1200         2591.90
 2         900              900                645         1764.00
 3        1400             1400                880         2576.00
 4        1500               58                920          172.54

City   Tankered amt   Tankering capacity
 1        150              860
 2        393              960
 3        900              900
```

图 24.5 AMPL 模型的输出

看起来奇怪的是, 图 24.5 的解要求在经停站 4(FOE) 按照每加仑 \$3.00 的价格购买 58 加仑燃油, 而不是简单地在经停站 1(DEN) 以 \$1.92 的价格多买, 而且那里还多加了 150 加仑的燃油. 不在经停站 1(DEN) 购买更多燃油的原因是, 所给出的最优解要在经停站 3(STL) 利用整个的加油容量 (=900 加仑). 事实上, 假如在 STL 的加油容量还能增大的话, 例如增大到 1 000 加仑, 则在 FOE 根本就不用购买.

思考题

1. 在图 24.5 中, 验证`Tankered amt`项下的值.
2. 给定 DEN, MCI, STL, FOE 的基准价格分别为 \$2.40, \$2.70, \$2.80, \$3.00, 用启

发式算法求机动加油策略.

3. 在第 1 题中, 若飞机的装油容量为 1 000 加仑, 并且每个经停站最多能提供 600 加仑, 求机动加油策略, 先用启发式算法, 再用线性规划求解.
4. 说明若在经停站 3(STL) 的装油容量从 900 加仑提高到 1 000 加仑, 则不用在经停站 4(FOE) 购买燃油, 并且在经停站 2(MCI) 的购买量可增加以满足航段 4 的燃料需求.

案例 2 心脏瓣膜的最优生产计划①

工具: LP

应用领域: 生物假体 (生产计划)

问题的描述

生物心脏瓣膜是用猪的心脏生产的生物假体, 用于人类心脏瓣膜移植. 人类需要修复的瓣膜有着不同的大小规格. 从供应方面来看, 猪的心脏不可能按照指定的大小规格来“生产”. 此外, 所制造的瓣膜的确切规格必须要在猪心脏经过加工以后才能确定. 这样, 某些需要的规格可能过多, 也可能出现短缺.

原材料猪心脏由若干家供应商提供, 规格有 6 到 8 种, 通常根据动物饲养的方式按照不同比例提供. 每次供货的规格分布用直方图来表示. 养猪专家与供应厂商合作, 尽量保证分布的稳定性. 按照这种方法, 生产商对每次供货的每种规格的瓣膜数量能够有一个合理可靠的估计. 因此, 如何选择多家供应商、合理选择供货的规格, 对于降低供应与需求之间的匹配不当是至关重要的.

LP 模型

令

$m =$ 瓣膜规格的数量

$n =$ 供应商的数量

$p_{ij} =$ 由供应商 j 提供的规格 i 的原料瓣膜的比例数, $0 < p_{ij} < 1$,

$$i = 1, 2, \cdots, m,\ j = 1, 2, \cdots, n,\ \sum_{i=1}^{m} p_{ij} = 1,\ j = 1, 2, \cdots, n$$

$c_i =$ 规格 i 的原料心脏的购买和加工费用, $i = 1, 2, \cdots, m$

$\bar{c}_j =$ 来自供应商 j 的心脏瓣膜的平均费用 $= \sum_{i=1}^{m} c_i p_{ij}, j = 1, 2, \cdots, n$

① 资料来源: S. S. Hilal and W. Erikson, “Matching Supplies to Save Lives: Linear Programming the Production of Heart Valves,” *Interfaces*, Vol.11, No.6, pp.48-55, 1981.

D_i = 对规格 i 瓣膜的平均月需求量

H_j = 供应商 j 可提供的最大月供应量, $j = 1, 2, \cdots, n$

L_j = 供应商 j 每月答应提供的最小供应量, $j = 1, 2, \cdots, n$

问题的变量可定义为

$$x_j = \text{由供应商 } j \text{ 提供的月供应量 (心脏原料的数量)}, j = 1, 2, \cdots, n$$

这个线性规划求每个供应商提供的产品量, 在需求与供应约束条件下使得购买和加工总费用达到最小.

$$\begin{aligned}
\min \quad & z = \sum_{j=1}^{n} \bar{c}_j x_j \\
\text{s.t.} \quad & \sum_{j=1}^{n} p_{ij} x_j \geqslant D_i, \quad i = 1, 2, \cdots, m \\
& L_j \leqslant x_j \leqslant H_j, \quad j = 1, 2, \cdots, n
\end{aligned}$$

要得出完全正确的方案, 变量 x_j 必须限制为整数值. 但是, 参数 p_{ij} 和 D_i 仅仅是估计值, 因此, 在这种情况下把连续解舍入到最相近的整数值可能是一个不错的近似. 特别是对于规模较大的问题, 我们更需要这样做, 因为整数限制条件可能导致计算上的困难.

AMPL 计算

虽然用 AMPL 计算 LP 是相当简单的, 但输入数据会有些烦琐. 为模型提供数据的一种方便的做法是通过电子表格. 文件 excelCase2.xls 为模型给出了所有的表格, AMPL 程序文件 amplCase2.txt 说明了如何从 Excel 表格输入涉及 8 种瓣膜规格和 12 个供应商的数据.①

结果分析

用 excelCase2.xls 中的数据计算 AMPL 模型的输出结果如图 24.6 所示. 从严格意义上, 这个解的结果并不能用来安排实际生产, 因为对心脏瓣膜 i 的需求 D_i 是基于期望值的计算. 因此解 $x_j, j = 1, 2, \cdots, n$ 将导致有些月份剩余, 而有些月份短缺.

① 从文件 amplCase2.txt 所用到的电子表格程序 excelCase2.xls 中读入矩阵形式的数据有一个要求, ODBC 处理程序要求 Excel 读入表中的列标题必须是字符串, 即纯粹数字标题是不可接受的. 为了解决这个限制, 要利用 Excel TEXT 函数把所有的列标题转换成字符串. 这样, 第一列标题就可以用公式 =TEXT(COLUMN(A1), "0") 替换, 将这个公式复制到后面的列上将会自动把数字代码转换为所要的字符串.

```
Cost = $  42210.82
solution:
 j     L[j]      x[j]      H[j]     reduced cost    Av. unit price
 1      0        0.0       500        2.39             14.22
 2      0        0.0       500        0.12             15.88
 3      0        0.0       400        5.22             15.12
 4      0      116.4       500        0.00             14.70
 5      0      300.0       300       -0.49             16.68
 6      0      500.0       500       -2.13             14.89
 7      0      250.5       600        0.00             18.12
 8      0      400.0       400       -6.22             16.61
 9      0      300.0       300       -4.20             17.19
10      0      357.4       500       -0.00             14.47
11      0      112.9       400        0.00             15.62
12      0      293.1       500        0.00             16.31

 i     D[i]  Surplus[i]    Dual value
 1     275       0.0        29.28
 2     310      28.9         0.00
 3     400       0.0        19.18
 4     320      88.1         0.00
 5     400       0.0        24.33
 6     350       0.0         8.55
 7     300       0.0        62.41
 8     130      28.2         0.00
```

图 24.6 瓣膜生产模型的输出

那么这个模型有多大用处呢? 实际上, 这些结果可以用来作计划. 具体来说, 所得到的解建议把供应商分成 3 类:

(1) 厂商 1, 2, 3 必须从供应商名单中删除, 因为 $x_1 = x_2 = x_3 = 0$.

(2) 供应商 5, 6, 8, 9 对于满足需求是关键的, 因为解要求这些供应商提供他们所能生产的所有心脏.

(3) 其他的供应商 (4, 7, 10, 11, 12) 从满足需求的角度来看, 他们的重要性 "一般", 因为他们的最大生产能力没有得到充分的利用.

所给出的建议方案进一步由图 24.6 中简约费用值得到支持. 供应商 9 可以把它的平均单价再提高 \$4.00, 但仍可能在最优解中, 而供应商 3 即使把平均单价降低 \$5.00 还是缺乏吸引力. 这个结果是对的, 虽然实际上排除掉的供应商 3 的平均单价是最低的 (= \$15.12), 而且 "明星" 供应商 9 的平均价格是最高的 (= \$17.19). 得出这一明显不符合直觉的结论的原因是, 该模型主要是需求驱动的, 即供应商 5,

6, 8, 9 比其他供应商提供了相对多的所需要的规格. 反过来对于供应商 1,2,3 也是对的. 这就意味着, 需求水平的变化可能导致选择不同的供应商组合, 其原因是, 在相对稳定的需求量估计下, 生产厂商与"明星"供应商们密切合作, 给他们提供营养和饲养条件的建议, 保证他们的瓣膜规格分布保持合理的稳定状态.

瓣膜规格 7 在所有的规格中似乎是最关键的, 因为它有最高的对偶价格 (= \$62.41), 是其他规格瓣膜对偶价格的两倍多. 这说明应该密切监督规格 7 瓣膜的库存量, 保持它的剩余库存尽量处于最低的水平. 另一方面, 规格 2, 4, 8 的瓣膜显现出过剩, 必须努力减少它们的库存量.

对模型计算结果的评述

这个模型的结果给出了一般性的计划指导, 而不是确定的生产安排方案, 在这种意义下, 所提出的 LP 模型有些"初级". 但是如原文所述, 从提出的计划中节省的经费是不少的. 通过从所有供应商中去掉一部分厂商, 以及找出"明星"供应商, 我们减少了库存, 节约了大量的经费. 这个计划同样还能降低以前不用模型结果时经常发生的缺货情况. 并且, 通过找出最优惠的供应厂家, 有可能让生产厂的养猪专家来培训供应厂家屠宰场的工人, 以提供加工质量高的心脏产品. 这又会使得生产过程效率更高.

思考题

1. 若 8 个规格的瓣膜需求量分别 300, 200, 350, 450, 500, 200, 250, 180, 求出新的解, 并与案例分析给出的解进行比较.
2. 假设供应商 3 已经签订了下一年供货 400 只心脏的协议, 并且通过特殊饲养能改变所提供心脏规格的比例, 那么供应商 3 应该提供的各规格产品理想的比例是多少?

案例 3 澳大利亚旅游委员会关于旅游产品交易会的会面安排问题①

工具: 指派模型, 启发式算法
应用领域: 旅游业
问题的描述

澳大利亚旅游委员会 (Australian Tourist Commission, ATC) 在全球各地组织旅游产品交易会, 为澳大利亚的旅游销售商提供一种与国际购买商交流的平台, 推销包括旅店、旅游线路、交通等旅游产品. 交易会期间, 销售商设立展台, 按照事先

① 资料来源: A. T. Ernst, R. G. J. Mills, and P. Welgama, "Scheduling Appointments at Trade Events for the Australian Tourist Commission," *Interfaces*, Vol.33, No.3, pp.12-23, 2003.

安排好的会面计划接待购买商. 因为每次交易会的时间段有限, 而且卖方和买方公司的数量非常大 (1997 年在墨尔本举办的交易会上, 吸引了 620 个销售方和 700 个购买方), 因此 ATC 希望在交易会之前精心安排好卖方-买方之间的会面, 使得最大程度地符合双方的兴趣. 基本思路是让双方的兴趣能够匹配, 使得展会期间有限的时间段得到最有效的利用.

分析

这个问题可以看成是一个三维的指派模型, 表示买方、卖方和事先确定的时间段. 对于有 m 个买方、n 个卖方、T 个时间段的交易展会, 定义

$$x_{ijt} = \begin{cases} 1, & \text{若买方 } i \text{ 在时间段 } t \text{ 与卖方 } j \text{ 见面} \\ 0, & \text{否则} \end{cases}$$

$$c_{ij} = \text{某个数, 代表买方 } i \text{ 与卖方 } j \text{ 双方的偏好或愿望}$$

相应的指派模型可以表示为

$$\max\ z = \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}\left(\sum_{t=1}^{T} x_{ijt}\right)$$

$$\begin{aligned}
\text{s.t.}\ & \sum_{i=1}^{m} x_{ijt} \leqslant 1, \quad j = 1,2,\cdots,n, \quad t = 1,2,\cdots,T \\
& \sum_{j=1}^{n} x_{ijt} \leqslant 1, \quad i = 1,2,\cdots,m, \quad t = 1,2,\cdots,T \\
& \sum_{t=1}^{T} x_{ijt} \leqslant 1, \quad i = 1,2,\cdots,m, \quad j = 1,2,\cdots,n \\
& x_{ijt} \in \{0,1\} \quad \text{对于任意的 } i,\ j,\ t
\end{aligned}$$

这个模型表达了一个指派模型的基本限制条件: 每个买方或卖方每次会面最多能见一个人, 每次特定的买卖双方会面最多发生一次. 在目标函数中, 系数 c_{ij} 表示买卖双方会面的愿望, 不依赖于每次会面, 因为我们假定买方和卖方对于会面时间并不关心.

那么如何来确定系数 c_{ij} 呢? 在买方和卖方到展会注册时, 每家卖方要向 ATC 提供一份最希望会面的买方的名单, 而买方也要提供一份类似的卖方的名单. 名单对最希望会面的对方打 1 分, 这个值越大意味着见面的愿望越小. 这些名单不需要对所有的公司全部都打分, 因为买卖双方可以自由表达与某些公司会面的兴趣, 而不是所有注册的公司. 例如, 在一份列有 100 个卖方的名单中, 一个买方可能只希望会见 10 个卖方, 在这种情况下, 对这些选择的卖方所表示的会见愿望就会是 $1, 2, \cdots, 10$.

从买卖双方名单生成的原始数据就可以代数表达为

$b_{ij}=$买方 i 对会见卖方 j 所给出的愿望分

$s_{ji}=$卖方 j 对会见买方 i 所给出的愿望分

$B=$所有买方给出的愿望的最大数

$S=$所有卖方给出的愿望的最大数

$\alpha=$买方愿望的相对权重 (用来计算愿望分值 c_{ij}), $0<\alpha<1$

$1-\alpha=$卖方愿望的相对权重

根据这些定义, 目标系数可以计算为

$$c_{ij}=\begin{cases}1+\alpha\left(\dfrac{B-b_{ij}}{B}\right)+(1-\alpha)\left(\dfrac{S-s_{ji}}{S}\right), & \text{如果 } b_{ij}\neq 0 \text{ 并且 } s_{ji}\neq 0\\ 1+\alpha\left(\dfrac{B-b_{ij}}{B}\right), & \text{如果 } b_{ij}\neq 0 \text{ 并且 } s_{ji}=0\\ 1+(1-\alpha)\left(\dfrac{S-s_{ji}}{S}\right), & \text{如果 } b_{ij}=0 \text{ 并且 } s_{ji}\neq 0\\ 0, & \text{如果 } b_{ij}=s_{ji}=0\end{cases}$$

这些公式背后的逻辑是, b_{ij} 的值越小意味着 $(B-b_{ij})$ 的值就越大, 因此给买方 i 和卖方 j 之间的会面要求所赋的分值就越高. 对卖方 j 要求与买方 i 会见给出的分值 $S-s_{ji}$ 也有类似的解释. 将这两个分数值分别用 B 和 S 相除归一化成 0 到 1 之间的值, 然后用 α 和 $1-\alpha$ 加权, 反映买方和卖方愿望的相对重要性, $0<\alpha<1$. 这样, α 的值小于 0.5 偏向于卖方的愿望. 注意到, $b_{ij}=0$ 和 $s_{ji}=0$ 表示买方 i 和卖方 j 之间不要求会面. c_{ij} 的前 3 个公式中出现的量 1 让它与不要求会面的情况 (即 $b_{ij}=s_{ji}=0$) 相比有相对更大的愿望. 原始分数的归一化保证了 $0\leqslant c_{ij}<2$.

输入数据的可靠性

现在, 一个关键的问题就是由买方和卖方给出的偏好数据的可靠性. 我们设计了一个采集偏好数据的工具, 保证满足下面的条件.

(1) 卖方和买方的名单只有在过了注册期限以后才确定下来.

(2) 只有注册的卖方和买方公司能参与这个过程.

(3) ATC 对参与公司的愿望偏好严格保密, 其他参与公司不能看到或修改这些名单.

在这些条件下, 建立一个交互式的互联网站, 让参与的公司方便地输入自己的会面愿望. 更重要的是, 网站的设计要保证输入数据的正确性. 例如系统要防止买方与同一个卖方公司见面多次, 反之亦然.

问题的求解

本案例给出的指派模型很直观，可以用任何商业软件包来求解. 文件 amplCase3a.txt 和文件 amplCase3b.txt 提供了两个求解该问题的 AMPL 程序模型. 这两个模型的数据用电子表格式给出 (文件 excelCase3.xls). 在第一个模型中, 可以用这个电子表格计算系数 c_{ij}, 然后用作输入数据. 在第二个模型中, 输入数据是原始偏好分数 b_{ij} 和 s_{ji}, 模型自己计算系数 c_{ij}. 第二个模型的优点是, 它能计算卖方和买方关于他们表达的愿望的满意度百分数.

图 24.7 给出了模型 amplCase3b.txt 用文件 excelCase3.xls 中的数据 (6 个买方、7 个卖方和 6 次会面) 计算的输出结果, 它给出了每次会面指派的卖方和买方, 还有每个买方和每个卖方对权系数 $\alpha = 0.5$ 的满意度百分数. 这些结果说明买卖双方的满意度都很高 (分别为 92% 和 86%). 若取 $\alpha < 0.5$ 的话, 卖方的满意度还会提高.

一些实用方面的考虑

为了让指派模型的解更加切合实际, 就必须要考虑连续两次会面之间的延误时间. 本质上, 一个买方公司一旦完成了一次会面, 很可能就会到下一个展位去会见另一家公司. 一个可行的安排就必须考虑到连续两次会面之间的转移时间. 下面的行走约束能得到这一结果:

$$x_{ijt} + \sum_{k \in J_i} x_{i,k,t+1} = 1, \quad i = 1, \cdots, m, \quad j = 1, \cdots, n, \quad t = 1, \cdots, T$$

集合 J_i 表示买方 i 不能按时到达时段 $t+1$ 安排的会面, 逻辑是, 若买方 i 与卖方 j 在时段 t 安排了一次会面 ($x_{ijt} = 1$), 则该同一个买方不能与未按时到达的卖方 k 安排下一个时段 $(t+1)$ 的会面 (即 $x_{i,k,t+1} = 0$). 可以通过取消一些会面时间结束 (如茶歇、午餐和关门) 前的一些时段 t 来减少这样的约束数量.

附加约束会大大增加计算上的难度. 事实上, 因为考虑到现有 IP 算法在计算上的局限性, 可能导致作为整数线性规划模型是不可解的. 这也是为什么我们需要启发式算法来求出该问题的一个“满意”解的原因.

用来求解这个新的限制模型的启发式算法如下.

对每个时段 t, 执行

(1)若买方 i 在时段 $t-1$ 的上一次会面地点使他不能够在时段 t 与卖方会面, 则设 $x_{ijt} = 0$.

(2) 若 i 和 j 之间的一次会面已经事先安排好了, 则设 $x_{ijt} = 0$.

(3) 求解所得的二维指派模型.

转向下一个时段 t

```
Optimal score =   50.87
Optimal assignments:
Session 1:
  Assign buyer 1 to seller 1
  Assign buyer 2 to seller 5
  Assign buyer 3 to seller 4
  Assign buyer 4 to seller 6
  Assign buyer 5 to seller 2
  Assign buyer 6 to seller 7
Session 2:
  Assign buyer 1 to seller 3
  Assign buyer 2 to seller 6
  Assign buyer 3 to seller 5
  Assign buyer 4 to seller 2
  Assign buyer 5 to seller 1
  Assign buyer 6 to seller 4
Session 3:
  Assign buyer 1 to seller 2
  Assign buyer 2 to seller 4
  Assign buyer 3 to seller 6
  Assign buyer 4 to seller 5
  Assign buyer 5 to seller 3
  Assign buyer 6 to seller 1
Session 4:
  Assign buyer 1 to seller 5
  Assign buyer 2 to seller 3
  Assign buyer 3 to seller 1
  Assign buyer 4 to seller 7
  Assign buyer 5 to seller 4
  Assign buyer 6 to seller 2
Session 5:
  Assign buyer 2 to seller 2
  Assign buyer 3 to seller 3
  Assign buyer 4 to seller 4
  Assign buyer 5 to seller 5
  Assign buyer 6 to seller 6
Session 6:
  Assign buyer 1 to seller 4
  Assign buyer 2 to seller 7
  Assign buyer 3 to seller 2
  Assign buyer 4 to seller 1
  Assign buyer 5 to seller 6
  Assign buyer 6 to seller 5
Buyers satisfaction: Average =  92
Buyer:       1    2    3    4    5    6
Percent:   100   86  100   80   86  100
Sellers satisfaction: Average =  86
Seller:      1    2    3    4    5    6    7
Percent:    83  100   60  100  100  100   60
```

图 24.7 指派模型的 AMPL 输出

衡量启发式算法解的质量, 可以通过将得到的目标值 (愿望度量) 与原来的指派模型 (没有行走约束) 的目标值进行比较. 所得的结果表明, 对于 5 次不同的旅游交易会, 这两种解的差异小于 10%, 说明启发式方法能给出可靠的解.

当然, 这样得到的解由于时段数有限不能保证满足所有的会面愿望. 有意思的是, 启发式算法建议的解表明, 解中包含至少 80% 的优先会面 (愿望值等于 1), 这个百分数随着表达分值 (分值越高, 会面愿望越低) 的增长几乎线性地降低.

思考题

1. 假设展台的位置决定不能安排某个买方公司与卖方公司 1 和 3 举行相继会面. 修改 AMPL 模型, 加入行走约束并求出最优解.
2. 对第 1 题用启发式算法求解, 并与最优解比较.

案例 4　节省联邦政府的旅费支出①

工具: 最短路径算法

应用领域: 商务旅行

问题的描述

美国联邦政府雇员需要参加发展会议以及培训课程. 当然, 选择举办会议和培训班的城市时并没有考虑到将要发生的旅费开支. 由于联邦政府雇员来自美国各地的机构, 举办城市的地点就会影响到旅费的支出, 旅费的总支出取决于参会的人数和他们的出发地城市.

美国行政服务局 (General Services Administration, GSA) 每年发布一项年度机票计划表, 政府和各家航空公司签订了合同. 该计划表给出了 48 个州大约 5 000 对城市之间的飞机票价, 还规定了所有主要城市的补助费标准, 对没有包括在名单中的城市, 也给出了日差旅补助费率. 对于自驾车参加会议的可以按照每英里得到补助费. 所有这些费率每年进行修改, 以反映生活费的增长率. 从一个地点到主办城市的旅费开支是参加人数、旅行费用以及主办城市补助费的直接函数.

我们的问题是, 给定全国各地的参加人数, 求一项活动的最优举办城市地点.

分析

求解的思路很简单: 举办城市必须有最低的旅费支出, 包括 (1) 交通费用最少; (2) 举办城市的补助费最少. 确定交通费需要知道参加会议的人员从哪里出发. 可以合理地假定, 离举办城市 100 英里以内的地点, 参加人员选择自驾车作为交通工具, 其他地方的人乘飞机来. 乘飞机的费用标准由飞到举办城市的沿着最经济路

① 资料来源: J. L. Huisingh, H. M Yamauchi, and R. Zimmerman, "Saving Federal Travel Dollars," *Interfaces*, Vol.31, No.5, pp.13-23, 2001.

线航行的合同机票总和确定. 为了找到这样的路线, 必须要知道参会人员从美国的什么地点出发. 每个这样的地点如果有机场和必要的会议设施的话, 都可能成为会议举办地的候选城市. 对于本案例, 我们找到了 261 个这样的地点, 有 4 640 条合同机场航线.

要在选出的 261 个地点和 4 640 条航线中确定最经济的路线并非易事, 因为一趟旅行可能涉及多个航段. Floyd 算法 (见 6.3.2 节) 对求这样的路径是最理想的. 两个地点之间的“距离”我们用政府提供的合同票价来表示. 根据合同, 往返机票的费用等于单程票价的两倍.

为了简化分析, 这项研究不允许在目的地租车自驾. 这条似乎不太合理的假设是因为, 举办地旅馆都在机场附近, 通常都有免费机场接送服务.

差旅补助费包括住宿、餐费以及零花费用. 参会者于会议开始前一天到达. 但是, 那些从 100 英里以内来开会的人, 在开会第一天早上到达. 所有会议代表在最后一天退出旅店. 在到达当天和离开当天, 政府规定只能按照全额补助费率的 75% 报销餐费和零用费.

数值实例

为了详细说明, 我们假定有 12 个举办城市的一个实例. 表 24.3 给出了这些城市之间允许的合同单程票价. 空白单元表示相应的城市对之间没有直接的航线.

表 24.3　以 12 个城市为例的单程机票价格 ($)

	SF	ORD	STL	LAX	TUL	DEN	DC	ATL	DAL	NY	MIA	SPI
SF				70		120			220			
ORD			99			140	150					
STL		99			95	110						78[a]
LAX	70					130						
TUL			95			105			100			
DEN	120	140	110	130	105							
DC		150						100	195	85		
ATL							100				125	
DAL	220				100		195					
NY							85				130	
MIA								125		130		
SPI			78[a]									

(a) 机票费 =$78. 距离 < 100 英里 (=86 英里). STL 与 SPI 之间采用自驾车方式.

表 24.4 给出了这 12 个城市一次会议的最高住宿费和补助费, 以及相应的参会人数. 会议期间为 4 天, 自驾车的标准里程补贴为每英里 $0.325(2000 年标准).

表 24.4 以 12 个城市为例的住宿费、补助费和参会人数

城 市	每晚住宿费 ($)	补助费 ($)	参会人数
SF	115.00	50.00	15
ORD	115.00	50.00	10
STL	85.00	48.00	8
LAX	120.00	55.00	18
TUL	70.00	35.00	5
DEN	90.00	40.00	9
DC	150.00	60.00	10
ATL	90.00	50.00	12
DAL	90.00	50.00	11
NY	190.00	60.00	12
MIA	120.00	50.00	8
SPI	60.00	35.00	2

求解的第一步是确定所有城市对之间的最经济票价, 这一步由 TORA 用 Floyd 最短路径算法 (6.3.2 节) 计算. 表 24.5 是该算法的计算结果, 空白单元对称地等于对角线以上的值, 前面说过, 这些值代表单程机票价格, 而且往返票价等于单程机票的两倍. Floyd 算法自动指定每对城市的旅行航段.

表 24.5 以 12 个城市为例的最低票价 ($)

	ORD	STL	LAX	TUL	DEN	DC	ATL	DAL	NY	MIA	SPI
SF	260	230	70	225	120	410	510	220	495	625	308
ORD		99	270	194	140	150	250	294	235	365	177
STL			240	95	110	249	349	195	334	464	28*
LAX				235	130	420	520	290	505	635	318
TUL					105	295	395	100	380	510	173
DEN						290	390	205	375	505	188
DC							100	195	85	215	327
ATL								295	185	125	427
DAL									280	410	273
NY										130	412
MIA											542

* 自驾车费用按 86 英里计算 (每英里 32.5 美分).

求解的最后一步是计算这次会议所有参会者的总费用, 假定会议在其中某个城市举行, 使得总费用最低的城市将被选作会议地举办地点.

为了说明计算过程, 假定 STL 是候选举办城市, 则相应的总费用计算如下:

$$
\begin{aligned}
\text{旅费} &= 2 \times (15 \times 230 + 10 \times 99 + 18 \times 240 + 5 \times 95 + 9 \times 110 + 10 \times 249 \\
&\quad + 12 \times 349 + 11 \times 195 + 12 \times 334 + 8 \times 464) + 2 \times (2 \times 86) \times 0.325 \\
&= \$53\ 647.80
\end{aligned}
$$

$$
\begin{aligned}
\text{住宿费} &= \$85 \times [(15+10+18+5+9+10+12+11+12+8)\times 4 + 2\times 3] \\
&= \$37\,910
\end{aligned}
$$

$$
\begin{aligned}
\text{补助费} &= \$48 \times [(15+10+18+5+9+10+12+11+12+8) \\
&\quad \times 4.5 + (2+8)\times 3.5] = \$25\,440
\end{aligned}
$$

注意到因为 SPI 离 STL 86 为英里 (< 100 英里), 从那里来的代表采取自驾车方式, 并于会议第一天早上抵达 STL. 因此, 他们的补助费按 $3\frac{1}{2}$ 天算, 住宿费只有 3 个晚上. 来自 STL 的参会者补助费按 $3\frac{1}{2}$ 天, 没有住宿费. 其他所有代表一天前抵达 STL, 他们的补助费按 $4\frac{1}{2}$ 天算, 有 4 个晚上的住宿.

对所有举办城市可以方便地用图 24.8 所示的电子表格来计算 (文件 excelCase4.xls, 所有的公式都加在每个单元格注释上). 计算结果说明, TUL 提供了最低的总费用 ($108 365), 接下来是 DEN($111 332), 然后是 STL($115 750).

													N
Lodging	115	115	85	120	70	90	150	90	90	190	120	60	
Per-diem	50	50	40	55	35	40	60	50	50	60	50	35	
# participants	15	10	8	18	5	9	10	12	11	12	8	2	120
Event days	4												

Air travel cost matrix(determined by Floyd's algorithm):

	SF	ORD	STL	LAX	TUL	DEN	DC	ATL	DA	NY	MIA	SPI	Travel	Lodging	Per-diem	Total	
SF		260	230	70	225	120	410	510	220	495	625	308	$64,202	$48,300	$26,250	$138,752	SF
ORD	260		99	270	194	140	150	250	294	235	365	177	$51,220	$50,600	$26,500	$128,320	ORD
STL	230	99		240	95	110	249	349	195	334	464	28	$53,640	$37,910	$25,448	$116,998	STL
LAX	70	270	240		235	130	420	520	290	505	635	318	$66,842	$48,960	$28,710	$144,512	LAX
TUL	225	194	95	235		105	295	395	100	380	510	173	$58,052	$32,200	$18,725	$108,977	TUL
DEN	120	140	110	130	105		290	390	205	375	505	188	$51,392	$39,960	$21,240	$112,592	DEN
DC	410	150	249	420	295	290		100	195	85	215	273	$55,836	$66,000	$31,800	$153,636	DC
ATL	510	250	349	520	395	390	100		295	185	125	427	$72,212	$38,880	$26,400	$137,492	ATL
DA	220	294	195	290	100	205	195	295		280	410	273	$56,082	$39,240	$26,450	$121,772	DA
NY	495	235	334	505	380	375	85	185	280		130	412	$69,652	$82,080	$31,680	$183,412	NY
MIA	625	365	464	635	510	505	215	125	410	130		542	$92,132	$53,760	$26,600	$172,492	MIA
SPI	308	177	28	318	173	188	273	427	273	412	542		$70,064	$28,320	$18,830	$117,214	SPI

Note: Travel cost between STL and SPI is based on mileage (=86 * 325=$28)

图 24.8 举办城市的费用比较 (文件 excelCase4.xls)

思考题

1. 已知 TUL 为所选择的举办城市, 该如何为 SF 的代表作预订呢?
2. 假定在允许路线上加入下列新的航线: ORD-NY ($150), LAX-MIA ($300), DEN-DAL ($110), ATL-NY ($150), 求最优举办城市.

案例 5 泰国海军运送新兵最优行船路线及人员指派问题①

工具: 运输模型, ILP

应用领域: 交通运输, 路线编排

① 资料来源: P. Choypeng, P. Puakpong, and R. Rosenthal, "Optimal Ship Routing and Personnel Assignment for Naval Recruitment in Thailand," *Interfaces*, Vol.16, No.4, pp.47-52, 1986.

问题描述

泰国海军每年征兵 4 次, 招募到的新兵到全国 34 个征兵中心报到, 然后由大巴车运送到 4 个海军分基地, 从那里, 新兵们再乘船被送到海军总基地. 运兵船的所有航程都从海军总基地出发, 再回到总基地. 分基地的码头对靠岸的船型可能有一定的限制, 表 24.6 汇总了 3 种运输船的情况, 表 24.7 给出了总基地 (0) 与分基地 (1,2,3,4) 之间的距离数据. 表中的矩阵是对称的, 距离单位为公里.

表 24.6 可运送新兵的船

船型	容量 (运送新兵数)	往返行程数	允许停靠的基地
1	100	2	1
2	200	10	1,2,3
3	600	2	1,2,3,4

表 24.7 总基地 (0) 和分基地 (1, 2, 3, 4) 之间的距离矩阵 (公里)

从 \ 到	0	1	2	3	4
0	0	310	510	540	660
1	310	0	150	190	225
2	510	150	0	25	90
3	540	190	25	0	70
4	660	225	90	70	0

每个分基地有容量限制, 但总体上, 这 4 个基地能够容纳所有的新兵. 1983 年夏季, 共有 2 929 名新兵从征兵中心被送到这 4 个分基地, 然后再运送到总基地.

关于运送新兵有两个问题:

(1) 如何把这些新兵从征兵中心用大巴车送到分基地?

(2) 如何把新兵从分基地用船运送到总基地?

模型的建立

求解这个问题可以分成两个独立的阶段.

(1) 第 1 阶段: 求新兵从征兵中心到分基地的最优分配方案.

(2) 第 2 阶段: 利用第 1 阶段的结果, 求分基地到总基地的最优运输方案.

第一个问题是一个从 34 个出发地 (征兵中心) 到 4 个分基地 (目的地) 的直接运输模型. 为了建模, 出发地点的“供应量”就是征兵中心征募的新兵数, 一个目的地点的“需求量”等于一个分基地可接收的新兵数. 在这个问题中, 总的供应量不超过总需求量. 单位运输费用等于一辆大巴车的运输费用除以车上的座位数. 可以把运输单位表示为一辆大巴车的装载量, 而不是一个个的新兵. 用一辆车上的座位数去除一个征兵中心的新兵数, 再向上取整数即可得到所需的车辆数. 例如, 500 名新兵从一个征兵中心乘坐 52 座的大巴车需要 10 辆车来运送. 这种情况下单位

运输费用就是每辆车一次运送的费用.

不论我们把一个新兵还是一次车载量作为运输单位, 这两种表示法都是某种程度的近似. 把一个新兵作为运输单位可能少算总的运输费用, 因为车辆坐不满的时候和满员时的费用相同. 另一方面, 用整车数表示运输单位可能高估新兵数和分基地的容量. 不论用哪种方式, 解的偏差都不会太明显, 因为这种偏差只是来自把坐不满的车辆作为满员车辆来处理. 在本问题中, 我们把每个新兵作为运输单位来处理.

令

$x_{ij}=$ 征兵中心 i 运送到分基地 j 的新兵数

$c_{ij}=$ 征兵中心 i 运送到分基地 j 的每名新兵的运输费用

$a_i=$ 征兵中心 i 的新兵数

$b_j=$ 分基地 j 的容量 (新兵数)

数学模型为

$$\min\ z=\sum_{i=1}^{m}\sum_{j=1}^{n}c_{ij}x_{ij}$$

$$\text{s.t.}\ \sum_{j=1}^{n}x_{ij}\geqslant a_i,\quad i=1,2,\cdots,m$$

$$\sum_{i=1}^{m}x_{ij}\leqslant b_j,\quad j=1,2,\cdots,n$$

$$x_{ij}\geqslant 0,\ \text{对于任意的}\ i,\ j$$

只要满足一个可接受的假设, 即 $\sum_{i=1}^{m}a_i\leqslant\sum_{j=1}^{n}b_j$, 这个模型就有可行 (整数) 解.

由于上述模型很直观, 不需要做进一步分析. 这里, 我们假定该模型可用相应的数据求解, 得到的最优解如表 24.8 所示.

表 24.8 用大巴车运送到分基地的新兵数

分基地	新兵数
1	475
2	659
3	672
4	1 123
总数	2 929

第 1 阶段得到的最优解作为第 2 阶段的输入数据, 求从 4 个分基地运送这些新兵到总基地的最优运输计划, 这个解必须提供运送路线和船型. 一趟运送从总基

地 (0) 出发, 分段航行到达一个或多个分基地 (1,2,3,4), 要考虑到表 24.6 给出的港口对 3 种船型所允许的靠岸条件. 例如, 船型 2 只能靠岸分基地 1, 2, 3.

建立这个模型的第一个任务是考虑到港口停泊的限制, 找出所有可能的航行路线. 船型 1 只能去到分基地 1, 因此相应的往返航线为 0-1-0. 对于船型 2, 允许靠岸的码头只能有分基地 1, 2, 3. 为了求解, 船型 2 的往返航线可能包括一个基地 (如 0-1-0)、两个基地 (如 0-1-2-0) 或 3 个基地 (如 0-1-2-3-0). 同样的思路用来分析船型 3 的航线, 可能包括从分基地 1 到分基地 4 的所有情形. 重要的是要知道, 涉及相同基地的两条往返航线不一定产生相同的总航行距离. 例如, 航线 0-1-3-4-0 和 0-1-4-3-0 涉及相同的分基地 (1,3,4), 但这两条航线的总距离是不一样的, 即

$$\text{距离}(0\text{-}1\text{-}3\text{-}4\text{-}0) = 310 + 190 + 70 + 660 = 1\ 230\ \text{km}$$

$$\text{距离}(0\text{-}1\text{-}4\text{-}3\text{-}0) = 310 + 225 + 70 + 540 = 1\ 145\ \text{km}$$

因此, 必须考虑所有往返航线的各种排列情况.

令

$J =$ 所有 (往返) 航线的集合

$A_i =$ 包含基地 i 的航线的集合, $i = 1,2,3,4$

$B_j =$ 航线 j 到达的基地的集合, $j \in J$

$C_k =$ 使用船型 k 的航线的集合, $k = 1,2,3$

$x_j =$ 使用航线 j 的次数, $j \in J$

$y_{ij} =$ 从基地 i 走航线 j 所运送的新兵数, $i = 1,2,3,4, j \in J$

$r_i =$ 基地 i 的新兵数 (由第一阶段求出), $i = 1,2,3,4$

$n_k =$ 船型 k 可用的船数, $k = 1,2,3$

$c_k =$ 船型 k 的运送容量, $k = 1,2,3$

$d_j =$ 航线 j 的总长度 (单位: 公里), $j \in J$

行船路径问题的模型为

$$\begin{aligned}
\min\ & z = \sum_{j\in J} d_j x_j \\
\text{s.t.}\ & \sum_{j\in A_i} y_{ij} \geqslant r_i,\quad i = 1,2,3,4 \\
& \sum_{i\in B_j} y_{ij} \leqslant c_k x_j,\quad j \in C_k,\ k = 1,2,3
\end{aligned}$$

$$\sum_{j\in C_k} x_j \leqslant n_k,\quad k=1,2,3$$

$$x_j \geqslant 0,\ \text{取整数},\ j\in J$$

$$y_{ij} \geqslant 0,\ \text{取整数},\ i=1,2,3,4,\ j\in J$$

第 1 个约束保证所有的新兵都被运送完, 第 2 个约束体现每种船型的容量限制. 第 3 个约束保证所用每种船型的往返航线不超过已有的船数.

AMPL 求解

文件 amplCase5.txt 提供了求解第二阶段问题的 AMPL 模型. 该模型对每个船型自动生成所有可能的航线及相应的距离. 这一模型大概最“绕弯”的地方就是数学模型中定义的 A_i 和 B_j 这两个集合的确定, 通过建立矩阵 a_{ij} (若基地 i 在航线 j 上, 则令 $a_{ij}=1$; 否则 $a_{ij}=0$.) 来反映不同航线的航段. 例如, 若航线 30 为 0-1-4-3-0, 则 $a_{1,30}=1$, $a_{2,30}=0$, $a_{3,30}=1$, $a_{4,30}=1$. 利用矩阵 $\|a_{ij}\|$, 在约束中进行下列替换就不用直接求集合 A_i 和 B_j 了:

$$\sum_{j\in A_i} y_{ij} = \sum_{j\in J} a_{ij}y_{ij},\quad i=1,2,3,4$$

$$\sum_{j\in B_j} y_{ij} = \sum_{i=1}^{4} a_{ij}y_{ij},\quad j\in C_k,\ k=1,2,3$$

图 24.9 表示了用前面给出的数据计算模型的输出. 图中列出了航线、相应的船型以及该航线航行的次数. 每条航线还指定了每个基地来的新兵数. 行列汇总确定了解的可行性, 说明了 (1) 满足船的承载能力; (2) 运送的新兵数至少等于问题数据所指定的数目 (例如在基地 1, 所得到的解保证可以运送 541 名新兵, 超过了实际的 475 名新兵).

所得到的解“放大了”运送的新兵数, 其原因是, 我们把第一个约束指定为不等式而不是等式约束, 然而这并不意味着解会使用过多的“资源”. 这个不等式约束只是允许运送新兵数“放大”, 以利用船上仍然可用的过剩的运载容量. 事实上可以明显看出, 图 24.9 中的行汇总数恰好等于船的运载容量乘上航行次数.

为了说明不等式不会造成使用不必要的过多资源, 图 24.10 给出了同一个模型用同样数据, 但把第一个不等式约束换成等式约束的输出结果. 在这种情况下, 列求和恰好等于每个基地的新兵数, 而行求和说明, 各条船上“用到的”装载容量可以小于总的可装载容量. 例如, 航线 0-1-0 从基地 1 带走了 390 名新兵, 这个数小于船型 2 的最大装载容量 $(=2\times 200=400)$. 这个事实说明, 要用两次航行, 这刚好就是图 24.9 的解所给出的结果.

Total Ship kilometers = 10515

Ship class	Capacity	Tour legs	Nbr of trips	Nbr drafttees carried from base 1	2	3	4	Totals
2	200	0-1-0	2	400	0	0	0	400
2	200	0-1-2-0	3	141	459	0	0	600
2	200	0-2-1-0	1	0	200	0	0	200
2	200	0-1-2-3-0	3	0	0	600	0	600
3	600	0-1-2-4-3-0	2	0	0	77	1123	1200
			Totals	541	659	677	1123	

Total number of roundtrips = 11

图24.9　带有不等式约束（$\sum_{j\in A_i} y_{ij} \geq r_i$）的AMPL输出结果

Total Ship kilometers = 10515

Ship class	Capacity	Tour legs	Nbr of trips	Nbr drafttees carried from base 1	2	3	4	Totals
2	200	0-1-0	2	334	0	0	0	334
2	200	0-1-2-0	1	0	200	0	0	200
2	200	0-2-1-0	3	141	459	0	0	600
2	200	0-1-2-3-0	3	0	0	595	0	595
3	600	0-1-2-4-3-0	2	0	0	77	1123	1200
			Totals	475	659	672	1123	

Total number of roundtrips = 11

图24.10　第一个约束为等式（$\sum_{j\in A_i} y_{ij} = r_i$）的AMPL输出结果

计算方面的考虑

虽然上述两个 (采用不等式约束和等式约束) 模型产生相同的最优解, 但这两个整数线性规划的计算过程却是不一样的. 在等式约束情况下, AMPL 生成了总共 34 290 个分支定界节点, 这比不等式约束情况的 20 690 个节点多了 70%[①]. 这似乎是, 更加受限制的解空间本来应该使得分支定界算法的收敛更快, 因为从某种意义上, 算法对可行整数空间上设定了更加紧的限制. 不幸的是, 由于分支定界算法的 (通常是怪异的) 行为跟问题有很大的关系, 所以我们并不建议采用这样的规则.

思考题

1. 考虑下列往返航线.
   船型 1: 0-1-0
   船型 2: 0-1-0, 0-3-2-0, 0-1-2-3-0, 0-2-3-0
   船型 3: 0-3-0, 0-1-4-3-2-0, 0-3-1-0, 0-4-2-1-0, 0-3-4-1-0
   (a) 假设只允许有这些航线, 建立相应的 $\|a_{ij}\|$ 矩阵.
   (b) 写出第二阶段的 ILP 模型.
2. 假设在一个港口的每次进港和离港, 都相应有一笔固定的费用 \$5 000. 进一步假定, 对于船型 1, 2, 3 的每公里运行费用分别是 \$100, \$120, \$150. 利用本案例分析的数据, 求出从 4 个分基地到总基地的最优运输计划.

案例 6 Mount Sinai 医院手术室的时间分配问题[②]

工具: ILP, GP
应用领域: 医疗卫生
问题的描述

本问题发生在加拿大. 加拿大的医疗保险是强制性的, 所有公民必须投保. 经费的使用根据保费额和税收情况由各省控制. 在这套系统下, 向医院预先支付一笔固定的预算, 每个省再根据一种服务费投入机制向医生追加支付. 当地政府控制医疗系统的规模, 对医院的支出加以严格的限制. 结果造成医院对医疗资源, 尤其是手术室的使用必须要进行有效的控制.

Mount Sinai 医院有 10 个安排值班的手术室, 为 5 个科室提供服务, 即外科、妇科、眼科、耳鼻喉科和口腔外科, 其中有 8 个主手术室和 2 个临时性门诊外科 (EOPS) 手术室. 按照当天手术室使用的小时数, 一个手术室可以“短时使用”, 也

① 按照 CPLEX 9.1.3 程序计算得出.
② 资料来源: J. T. Blake and J. Donald, "Mount Sinai Hospital Uses Integer Programming to Allocate Operating Room Time," *Interfaces*, Vol.32, No.2, pp.63-73, 2002.

可以"长时使用". 由于加拿大医疗体制的公费医疗性质, 所有的外科手术只能安排在星期一到星期五之间. 表 24.9 汇总了不同类型手术室每天的可用情况, 表 24.10 给出了每周手术室时间的需求情况, 其中的可允许的分配不足的小时数是可以拒绝一个科室的最多小时数 (相对于其一周手术室时间请求).

表 24.9 Mt. Sinai 医院手术室可用时间

星期	可用小时数			
	主手术室"短时使用"	主手术室"长时使用"	EOPS"短时使用"	EOPS"长时使用"
星期一	08:00–15:30	08:00–17:00	08:00–15:30	08:00–16:00
星期二	08:00–15:30	08:00–17:00	08:00–15:30	08:00–16:00
星期三	08:00–15:30	08:00–17:00	08:00–15:30	08:00–16:00
星期四	08:00–15:30	08:00–17:00	08:00–15:30	08:00–16:00
星期五	09:00–15:30	09:00–17:00	09:00–15:30	09:00–16:00
手术室数量	4	4	1	1

表 24.10 手术室时间的每周需求小时数

科室	每周需求小时	可允许的分配不足的小时数
外科	189.0	10.0
妇科	117.4	10.0
眼科	39.4	10.0
口腔外科	19.9	10.0
耳鼻喉科	26.3	10.0
急诊科	5.4	3.0

这项研究的目标是为每天手术室安排确定一个合理均衡的计划, 实现现有手术室资源的最佳利用.

数学模型

对这个问题我们所能做的, 是要制订一个手术室的日安排方案, 最大限度地满足不同科室每周所需要的使用小时数. 换言之, 我们要对每个科室设定它的需求小时数作为目标, 并尽量满足要求. 模型的目标就是使得离每周需求小时的偏差达到最小.

令

x_{ijk} = 第 k 天分配给 j 科室第 i 种手术室的数量

d_{ik} = 第 k 天第 i 种手术室的时间长度 (可用小时数)

a_{ik} = 第 k 天第 i 种手术室的数量

h_j = j 科室请求使用手术室的 (理想) 小时数

u_j^- = j 科室允许分配不足的最大小时数

本问题涉及 6 个科室和 4 种类型的手术室, 因此 $i = 1,2,3,4$, $j = 1,2,\cdots,6$. 对每周 5 个工作日, 下标 k 取值从 1 到 5.

下面的整数目标规划模型表达了 Mount Sinai 医院的手术室安排调度问题:

$$\min \ z = \sum_{j=1}^{6} \left(\frac{1}{h_j}\right) s_j^-$$

$$\text{s.t.} \ \sum_{i=1}^{4}\sum_{k=1}^{5} d_{ik}x_{ijk} + s_j^- - s_j^+ = h_j, \ \text{对于任意的 } j \tag{1}$$

$$\sum_{j=1}^{6} x_{ijk} \leqslant a_{ik}, \ \text{对于任意的 } i, \ k \tag{2}$$

$$0 \leqslant s_j^- \leqslant u_j^-, \ \text{对于任意的 } j \tag{3}$$

$$x_{ijk} \geqslant 0 \ \text{且为整数}, \ \text{对于任意的 } i, \ j, \ k \tag{4}$$

$$s_j^-, s_j^+ \geqslant 0, \ \text{对于任意的 } j \tag{5}$$

这个模型的逻辑是, 我们不一定能满足所有科室 j, $j = 1,2,\cdots,6$ 所要求的手术室的理想小时数 h_j. 因此模型的目标是求一个手术室的安排方案, 使得不同科室可能“未能分配到的”手术室数达到最小. 为了达到这个目标, 约束 (1) 中的非负变量 s_j^- 和 s_j^+ 分别表示相对于要求的 h_j 分配不足的和分配过多的手术室. 比值 $\frac{s_j^-}{h_j}$ 度量对科室 j 分配不足的相对量, 约束 (2) 表示可用手术室的上限, 而约束 (3) 用来限制一个科室未满足的量, 上限值 u_j^- 由用户指定.

模型输出结果

文件 amplCase6.txt 给出了本问题的 AMPL 模型, 图 24.11 是用问题陈述中给出的数据所计算的解, 这个解表明所有的目标都得到了满足 ($z = 0$), 并给出了一周工作日期间 (星期一到星期五) 各科室详细的手术室 (按类型) 分配方案. 的确, 图 24.11 下方给出的各科室汇总情况表明, (6 个中) 有 5 个科室的请求都得到了过于充足的满足, 这刚好是因为这一周有足够的资源, 而模型并不对各科室多分配的手术室时间求最小. 实际上, 让现在这个模型去消除多余的分配小时并没有意义, 因为有手术室, 而且可以分派给不同的科室. 实质上, 我们主要关心的是当资源不能满足所有需求时分配不足的情况.

计算过程

在这个模型中, 变量 x_{ijk} 表示分配的手术室数, 因此必须为整数, 这就带来了一个我们常遇到的整数规划的计算问题. 我们的 AMPL 模型能够用问题描述中给出的数据迅速地计算出结果, 但是, 当我们把表示请求使用手术室的时间数 h_j 的数据稍微做些调整 (让其他数据保持不变) 时, 整个计算过程就完全变了. 首先, 计

算时间持续了一个多小时 (而初始的数据只需要几分钟), 而且在搜索了 4 500 多万个分支定界节点以后, 还不能得出一个可行解, 更不要说求出最优解了. 这种现象往往发生在当需求量超过供给能力的情况下. 实际上, 整数线性规划的这种计算行为是不可预测的, 因为当我们只要把目标函数改成求 s_j^- 不加权的和的最小值, 所有前面解不出来的情况就马上能求解了. 本案例后面的习题概述了这些计算过程的问题.

```
z = 0.00
Weekly Time Allocation:
Mon:
  Gynecology: 39.0 hrs
     4 room(s) type Main_L
     1 room(s) type Main_S
  Ophthalmology: 17.0 hrs
     1 room(s) type Main_S
     1 room(s) type EOPS_S
  Oral_surgery: 16.5 hrs
     1 room(s) type Main_S
     1 room(s) type EOPS_L
  Otolaryngology:  9.0 hrs
     1 room(s) type Main_S
Tue:
  Surgery: 17.0 hrs
     1 room(s) type Main_S
     1 room(s) type EOPS_S
  Gynecology: 39.0 hrs
     4 room(s) type Main_L
     1 room(s) type Main_S
  Oral_surgery: 7.5 hrs
     1 room(s) type EOPS_L
  Otolaryngology: 18.0 hrs
     2 room(s) type Main_S
Wed:
  Surgery: 66.5 hrs
     3 room(s) type Main_L
     4 room(s) type Main_S
     1 room(s) type EOPS_S
  Ophthalmology: 15.0 hrs
     1 room(s) type Main_L
     1 room(s) type EOPS_L
Thu:
  Surgery: 72.5 hrs
     4 room(s) type Main_L
     3 room(s) type Main_S
     1 room(s) type EOPS_L
     1 room(s) type EOPS_S
  Ophthalmology:  9.0 hrs
     1 room(s) type Main_S
Fri:
  Surgery: 34.0 hrs
     3 room(s) type Main_S
     1 room(s) type EOPS_S
  Gynecology: 39.0 hrs
     4 room(s) type Main_L
     1 room(s) type Main_S
  Emergency: 6.5 hrs
     1 room(s) type EOPS_L
Departmental summary:
  Surgery allocated 190.0 hrs (101%)
  Gynecology allocated 117.0 hrs (100%)
  Ophthalmology allocated 41.0 hrs (104%)
  Oral_surgery allocated 24.0 hrs (121%)
  Otolaryngology allocated 27.0 hrs (103%)
  Emergency allocated 6.5 hrs (120%)
```

图 24.11 Mt. Sinai 医院模型的输出结果

我们怎样做才能克服这些问题呢? 首先能想到的是, 能不能先把整数要求去掉, 然后再把得到的线性规划解进行凑整还原. 这种方法对本案例未必奏效, 因为很有可能得不出可行解. 假定我们有了医院手术室的具体数目, 很有可能各种试错凑整的解会达到现有手术室数的上限, 这就意味着没有任何一个方案能得出整数解来.

提高整数规划模型计算成功率的一种方法是, 通过考虑其他资源量来限制变量 x_{ijk} 的可行的取值范围. 例如, 某一天医院只有两个牙科医生的话, 这一天最多有两个 (任何类型的) 手术室可以分配给该科室. 设定这样的边界可能会有效地保证得到最优整数解. 如果不能满足这个要求, 唯一可以考虑的方法就是为这类问题设计一个启发式算法.

思考题

1. 在 amplCase6.txt 中, 解决如下问题.
   (a) 把 h_j 现在的值替换成
   ```
   param h:= Surgery 190  Gynecology 122  Ophthalmology 41.4
             Oral_surgery 20.9 Otolaryngology 25.3 Emergency 6.0;
   ```
   用新的数据执行该模型. 你会注意到, 模型运算一个多小时仍找不到一个可行解.
   (b) 利用模型原来的数据, 改变参数 a, 使得本周的所有 5 个工作日里 Main_L 只有 3 个手术室可用. 如 (a) 一样, AMPL 得不出一个解.
   (c) 在 (a) 和 (b) 两种情况下, 把目标函数都换成
   ```
   minimize z: sum{j in departments} (sMinus[j])
   ```
   求 s_j^- 的最小和. 然后执行该模型, 你会注意到, 对这两种情况能马上找到最优解.

案例 7 PFG 建材玻璃公司的拖车有效荷载优化问题①

工具: ILP, 静态受力 (自由体受力图) 计算

应用领域: 运输和分销

问题的描述

PFG 是一家南非的玻璃制造公司, 该公司用一种 (5 个轮子的) 特殊装备的拖车把平板玻璃包装箱运送到顾客那里. 包装箱的大小和重量通常都不同, 根据收到的订单, 拖车可能装载不同的玻璃箱. 图 24.12 是一个典型的车头和拖车装置, 点 A, B, C 是车轴的位置. 南非政府对轴重的最大上限有明确的规定, 而包装箱放在拖车上的位置对决定这些轴重至关重要. 把比较重的玻璃箱放在拖车的后部, 会增

① 资料来源: H. Taha, "An Alternative Model for Optimizing Payloads of Building Glass at BFG," *South African Journal of Industrial Engineering*, Vol.15, No.1, pp.31-43, 2003.

加车轴 A 的载重量, 而把它们放在拖车的前部, 就会把负载转移到车轴 B 和 C. 我们的目标是, 在不超过轴重上限的情况下, 我们要计算出拖车板上玻璃箱放置的具体位置顺序. 当然, 如果不考虑玻璃箱的摆放位置, 这种顺序的组合就可能超过轴重限制. 在这种情况下, 所得到的解就应该提供关于每个车轴超重的信息, 以便我们能够对顺序组合进行重新排列, 从而得出新的解. 重复这一过程, 直到所有轴重限制得到满足.

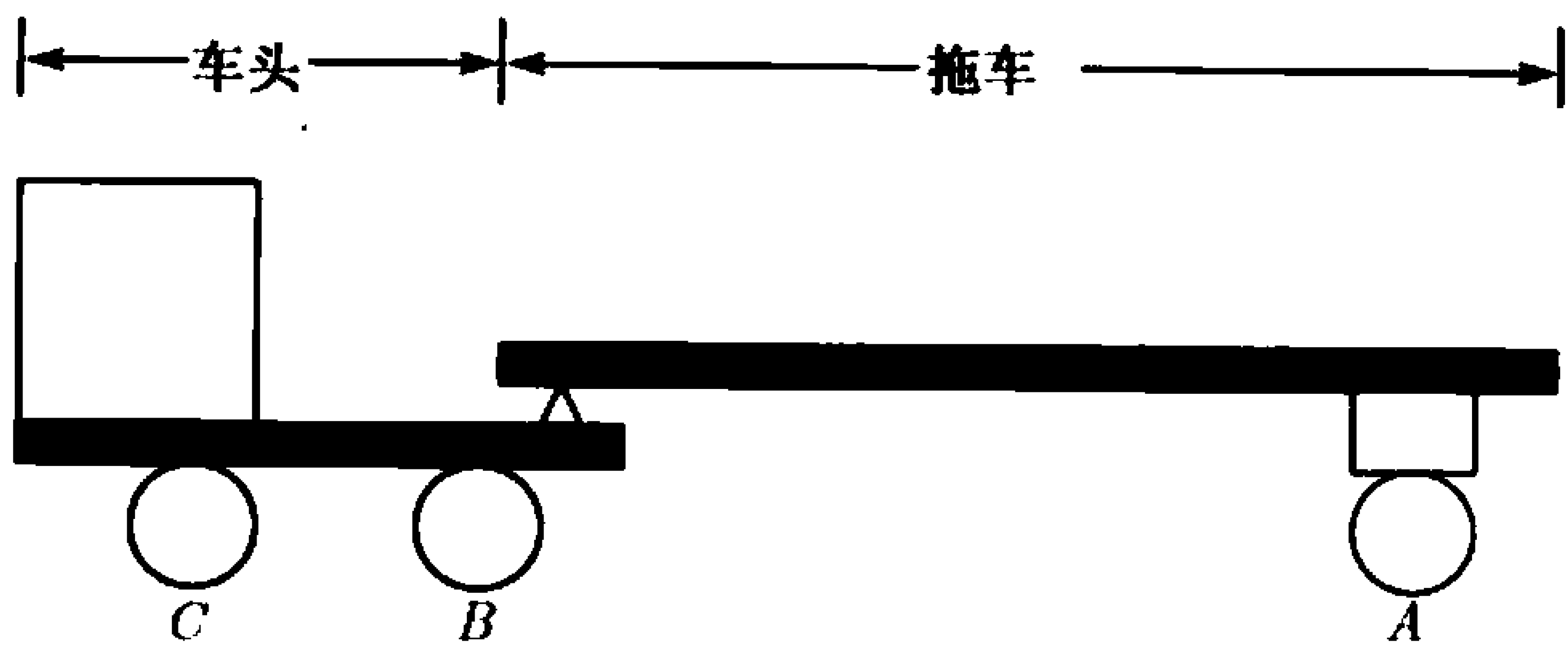


图 24.12 在车头–拖车装置中, 轴重的 A, B, C 点

根据这个问题的性质, 我们并不需要寻求这 3 个车轴上每个轴重的最小值, 因为减轻一个车轴的重量自动地就会增加另一个轴上的重量. 出于这个原因, 现实的目标是求出轴重的某种 "折中的" 分布. 在我们对问题详细分析以后, 就能够更明白这一点.

图 24.13 给出了拖车的一个示意性的俯视图. 整个拖车长度被分成若干区域, 其区域编号和边界大小依照玻璃箱的混装方式有所不同. 拖车的宽度从左到右对称地分成两边, 每边具有同样多的 "位置". 一个区域的一个位置用来放置一箱玻璃.

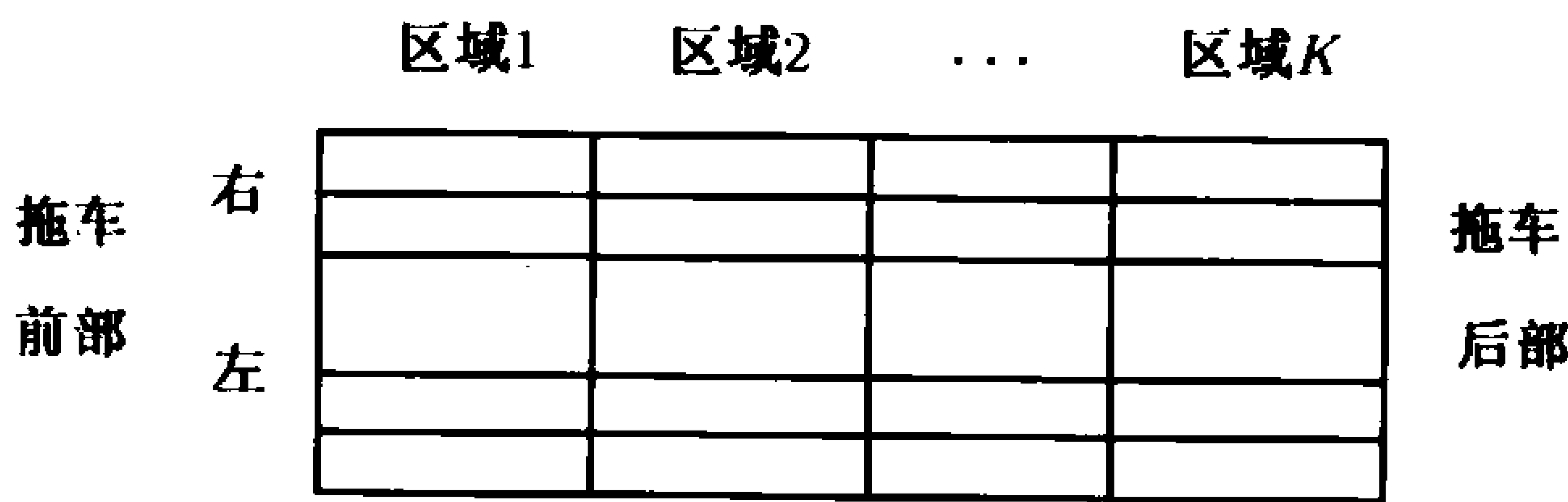


图 24.13 拖车位置的俯视示意图

轴重的计算

可以用下面的静态平衡方程来求出轴重, 先求拖车轴重, 然后再求车头的轴重.

(1) 合力 =0. (2) 合力矩 =0.

图 24.14 是拖车的自由体受力图. 为简单起见, 我们假定拖车有 3 个区域, 但分析方法可用于任意多个区域的情形. 对拖车上施加的力 (按千克) 为

$$f_k = \text{区域 } k \text{ 内玻璃箱重量}, k = 1, 2, 3$$

f_T = 拖车重量

R = 第 5 个轮子上的反作用力

f_A = 后轴处的反作用力

图中的不同距离 (米) 可由拖车的几何尺寸得出.

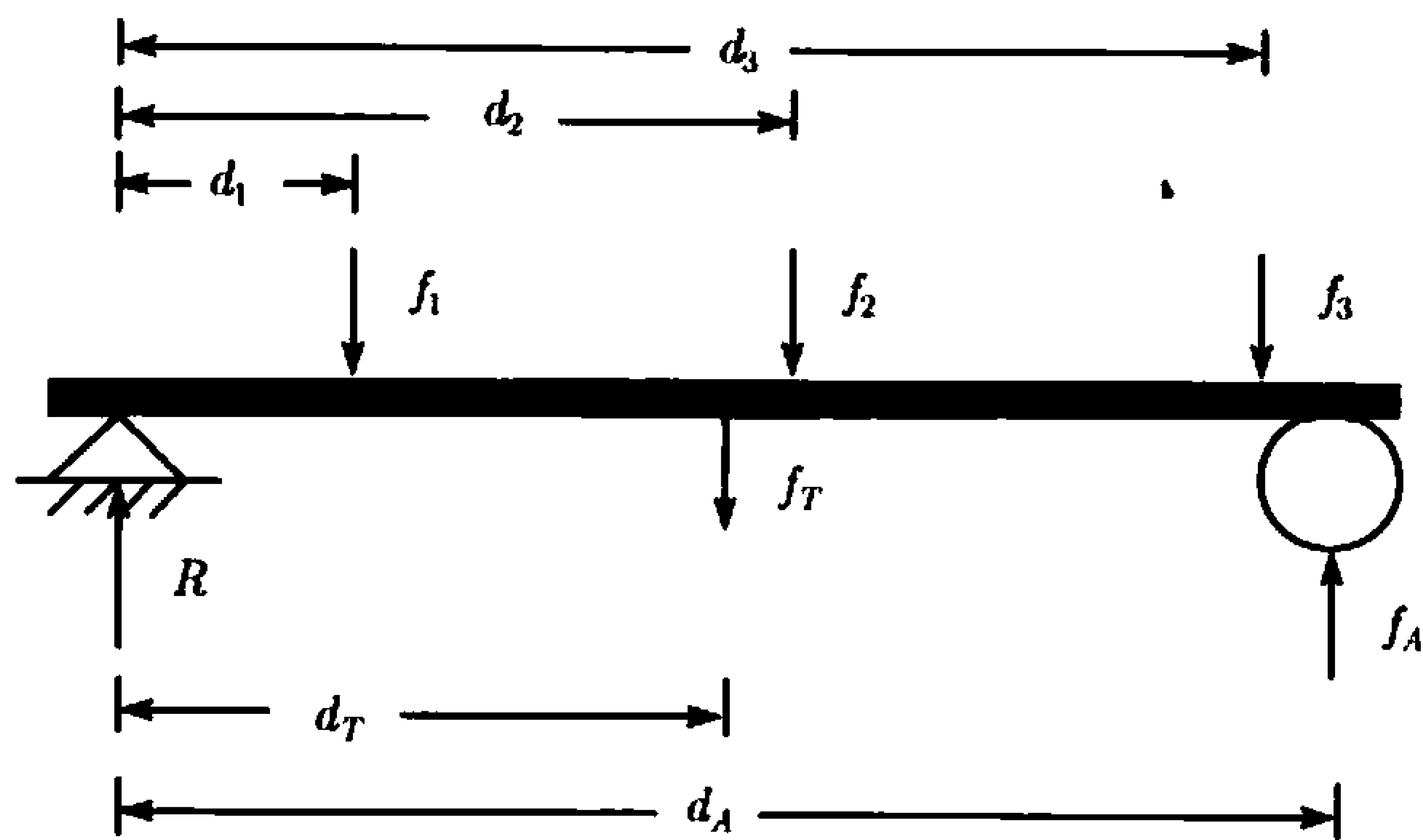


图 24.14　拖车自由体受力图

拖车的静态平衡方程为

$$f_1 + f_2 + f_3 + f_T = R + f_A$$

$$f_1 d_1 + f_2 d_2 + f_3 d_3 + f_T d_T - f_A d_A = 0$$

我们有两个未知变量的两个方程, 得到的解为

$$f_A = \frac{1}{d_A}(f_1 d_1 + f_2 d_2 + f_3 d_3 + f_T d_T) \tag{1}$$

$$R = f_1 + f_2 + f_3 + f_T - f_A \tag{2}$$

接下来, 我们对图 24.15 中车头的自由体受力图进行类似的分析. 车头的重量用 f_H 表示, 拖车受力对车头的作用等于上面方程 (2) 计算出的反作用力 R. 相应的平衡方程为

$$f_B + f_C = R + f_H$$

$$f_H d_H + R d_R - f_B d_B = 0$$

解出 f_B 和 f_C, 我们有

$$f_B = \frac{1}{d_B}(f_H d_H + R d_R) \tag{3}$$

$$f_C = R + f_H - f_B \tag{4}$$

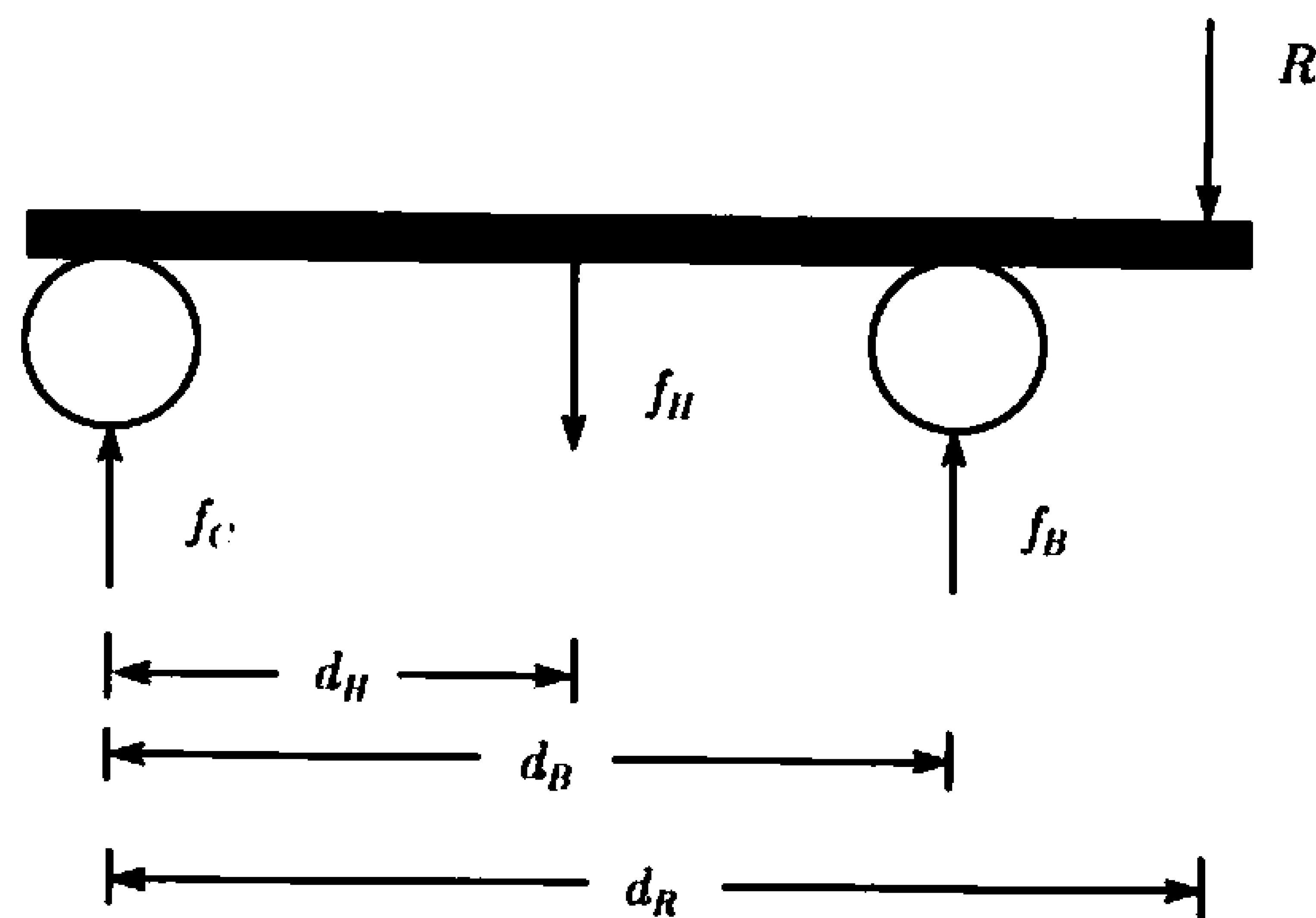


图 24.15 车头自由体受力图

方程 (1), (3), (4) 给出了作用在车头–拖车装置上车轴 A, B, C 上的反作用力的解 f_A, f_B, f_C. 令 w_A, w_B, w_C 表示车轴本身的固定负载, 则车轴上的净重量 (千克) 为

$$W_A = f_A + w_A \tag{5}$$

$$W_B = f_B + w_B \tag{6}$$

$$W_C = f_C + w_C \tag{7}$$

这些都是用于遵守政府规定的重量值.

数学模型

考虑有 I 种玻璃箱按某种顺序混合放在 K 个区域, 每个区域有 J 个位置的情形. 方程 (5),(6),(7) 给出的轴重可以用原始作用力 f_H, f_T, f_k, $k = 1, 2, \cdots, K$ 表示为

$$W_A = w_A + \frac{1}{d_A}\sum_{k=1}^{K} f_k d_k + f_T d_T \tag{8}$$

$$W_B = w_B + \frac{1}{d_B}(f_H d_H + R d_R) \tag{9}$$

$$W_C = w_C + R\left(1 - \frac{d_R}{d_B}\right) + f_H\left(1 - \frac{d_H}{d_B}\right) \tag{10}$$

其中

$$R = \sum_{k=1}^{K} f_k\left(1 - \frac{d_k}{d_A}\right) - \frac{1}{d_A}(f_T d_T)$$

定义

$$x_{ijk} = \begin{cases} 1, & \text{若将第 } i \text{ 种玻璃箱放在区域 } k \text{ 的 } j \text{ 位置} \\ 0, & \text{否则} \end{cases}$$

$$w_i = \text{每箱第 } i \text{ 种玻璃的重量}$$

放在拖车 k 区域的玻璃重量可用 x_{ijk} 计算为

$$f_k = \sum_{i=1}^{I} w_i \left(\sum_{j=1}^{J} x_{ijk} \right), \quad k = 1, 2, \cdots, K$$

(8), (9), (10) 中所有其他元素都是已知的.

在用 x_{ijk} 表示了轴重后, 现在就能构造最优化模型了. 该模型的主要目的是把玻璃箱摆放在特定位置上, 使得总的轴重 W_A, W_B, W_C 保持在政府规定的上限之内. 我们不能简单地在模型中把这些上限作为 $W_A \leqslant L_A$, $W_B \leqslant L_B$, $W_C \leqslant L_C$ 形式的约束, 因为这样做问题不一定有可行解. 在理想情况下, 我们希望求出 x_{ijk} 的值, 尽量地减小每一个轴重. 如果得出的最小轴重量满足政府规定, 则这个解就是可行的, 否则, 就必须要减少混装箱数, 并再找出一个新解. 求这种"理想化"解的困难是, 根据问题的特点, 不可能让每个轴重都达到最小, 因为如前所述, 这些轴重并不是独立的, 降低一个车轴的负载, 必然会增加另一个车轴的载重量.

一种接近于使得每个轴重都达到最小的模型是对**最大轴重求最小**, 这一模型的优点是集中考虑所有 3 个轴重的最极端情况. 目标函数的数学表示为

$$\min \ z = \max\{W_A, W_B, W_C\}$$

可以用下面的标准代换对这个函数进行线性化. 令

$$y = \max\{W_A, W_B, W_C\}$$

则这个目标函数可表示为

$$\begin{aligned} \min \ & z = y \\ \text{s.t.} \ & W_A \leqslant y \\ & W_B \leqslant y \\ & W_C \leqslant y \end{aligned}$$

现在来建立该模型的约束条件. 这些约束条件如下.

(1) 每个位置最多允许装一个玻璃箱.

(2) 不能超过区域重量上限.

(3) 拖车板上的装箱 (按类型) 数不超过最大混装数.

(4) 轴重不超过政府规定.

(5) 每个区域的玻璃箱尽量往中心放, 以防止拖车倾斜.

前 4 个约束是基本条件, 必须包含在模型里. 约束 (5) 虽然放在模型中, 但不重要, 因为区域 k 中的玻璃总重量是由 f_k 表示, 它并不是区域宽度的函数. 因此不论解如何, 实际的区域装载量会自动让玻璃箱往中心移动.

定义

$L_k =$ 区域 k 中玻璃载重量上限, $k = 1,2,\cdots,K$

$L_A =$ 轴 A(拖车后轴) 上的允许重量上限

$L_B =$ 轴 B(车头后轴) 上的允许重量上限

$L_C =$ 轴 C(车头前轴) 上的允许重量上限

约束 (1) 到 (5) 用数学公式表示为

$$\sum_{i=1}^{I} x_{ijk} \leqslant 1, \quad j = 1,2,\cdots,J; \quad k = 1,2,\cdots,K \tag{1}$$

$$\sum_{i=1}^{I} w_i \sum_{j=1}^{J} x_{ijk} \leqslant L_k, \quad k = 1,2,\cdots,K \tag{2}$$

$$\sum_{j=1}^{I} \sum_{k=1}^{K} x_{ijk} \leqslant N_i, i = 1,2,\cdots,I \tag{3}$$

$$W_A \leqslant L_A \tag{4a}$$

$$W_B \leqslant L_B \tag{4b}$$

$$W_C \leqslant L_C \tag{4c}$$

$$\sum_{i=1}^{I} x_{ijk} \leqslant \sum_{i=1}^{I} x_{i,j+1,k}, \ j = 1,2,\cdots,\frac{J}{2}-1; \ k = 1,2,\cdots,K \tag{5a}$$

$$\sum_{i=1}^{I} x_{ijk} \geqslant \sum_{i=1}^{I} x_{i,j+1,k}, \ j = \frac{J}{2},\frac{J}{2}+1,\cdots,J-1; \ k = 1,2,\cdots,K \tag{5b}$$

I 和 J 如前面定义的, 表示玻璃种类数和每个区域的位置数.

约束 (3) 没有设成等式约束, 是为了必要时可以让模型装载量少于订货量, 以满足由约束 (4) 所给出的规定上限. 但是由于模型的目标是要让最大轴重 y 达到最小, 而约束 (3) 是一个不等式, 因此显然最优解就是不让拖车板上装载任何玻璃箱. 为了解决这个问题, 我们把目标函数改成在不违反规定的前提下装载尽可能多的玻璃箱.

定义非负变量 s_k, $k = 1,2,\cdots,K$, 使得

$$\sum_{i=1}^{I} w_i \sum_{j=1}^{J} x_{ijk} + s_k = L_k, \quad k = 1,2,\cdots,K \tag{2a}$$

我们的思路是让这个松弛变量 $s_k, k = 1, 2, \cdots, K$ 能取最小值, 这样反过来也就要求装载由约束 (3) 所允许的最多的箱数. 现在的情况是, 我们需要求目标函数所定义的 y, 以及所有 $s_1, s_2, \cdots, s_K$ 的最小值. 为了达到这个目的, 一种合理的方法是把目标函数定义为

$$\min \ z = y + \sum_{k=1}^{K} s_k$$

这个新的目标函数的所有部分的单位都是千克.

AMPL 的求解策略

对上述给出的情形, 我们编制了一个 AMPL 模型. 不幸的是, 如许多典型的线性整数规划一样, 该模型运行了 4 个多小时也没能算出一个可行解. 这个"令人沮丧"的结果使我们不得不对数学模型进行修改. 去掉所有的约束 (4), 并采用了下面的求解策略.

第 1 步 运行这个 AMPL 模型. 如果最优解满足所有的轴重规定限制, 停止; 否则, 转到第 2 步.

第 2 步 去掉订货箱数最小的订单. 回到第 1 步.

这里的思路是, 从混装方案中去掉一些订货箱数, 最终达到某一点, 使得所有的规定限制得到满足. 在这种情况下, 去掉订货箱数最小的订单是合理的.

所提出的这个策略假定, 顾客必须一次得到所有的订货量. 还有一种方法是, 一次运送部分订单量, 以最大限度地利用拖车装载容量, 这样做显然更有利. 这种方法只需要把现在的策略改成把订货量一次减小一箱 (代替去掉整个订单量).

文件 amplCase7.txt 给出了实现所提出策略的程序. 这个程序无需另外解释, 因为它使用了数学模型中的符号. 该模型的`solve`程序段能够对得出解的可行性进行检查. 假如不满足规定的话, 它就会自动去掉最小订货量的订单, 并重新对模型进行优化. 继续这一过程, 直到实现可行性.

图 24.16 给出了针对某些特定的输入数据所得出的 AMPL 输出结果[①]. 输出部分表明, 第一个解不满足规定上限. 去掉第 3 号订单并重新优化后, 实现了可行性, 计算过程终止.

思考题

1. 对原先的问题建立 AMPL 模型, 其中所有的规定上限都用显式约束表示, 试着算出一个解.
2. 修改 amplCase7.txt 文件, 允许部分订货量. 所得到的解和图 24.16 的解有何不同?

① 用 CPLEX 9.1.3 计算, 大约 20 分钟得出了这个结果, 生成了 4 471 785 个 B & B 节点.

```
INPUT DATA:
Nbr of glass types = 4
    Packs of type 1 = 12
    Packs of type 2 = 6
    Packs of type 3 = 4
    Packs of type 4 = 7
Nbr of slots per zone = 10
Nbr of zones = 3

Weight limit for zone 1 = 10000 kg
Weight limit for zone 2 = 10000 kg
Weight limit for zone 3 = 10000 kg

Trailer axle weight = 3000 kg
Hauler rear axle weight = 1500 kg
Hauler front axle weight limit = 2000 kg

Total trailer axle weight limit = 24000 kg
Total hauler rear axle weight limit = 18000 kg
Total hauler front axle weight limit = 8000 kg

OUTPUT RESULTS:
Zone 1 weight = 9982 kg
Zone 2 weight = 9014 kg
Zone 3 weight = 6300 kg
Total weight at trailer axle = 25277.4 kg
Total weight at hauler rear axle = 15915.8 kg
Total weight at hauler front axle = 5902.9 kg
INFEASIBLE SOLUTION: Total trailer axle weight exceeds limit by 1277.4 kg

Packs of type 2 in zone 1 = 4
Packs of type 3 in zone 1 = 1
Packs of type 4 in zone 1 = 5
Packs of type 1 in zone 2 = 3
Packs of type 2 in zone 2 = 2
Packs of type 3 in zone 2 = 3
Packs of type 4 in zone 2 = 2
Packs of type 1 in zone 3 = 9

Searching for a feasible solution...attempt 1: Order nbr 3 removed
Zone 1 weight = 9930 kg
Zone 2 weight = 7866 kg
Zone 3 weight = 3500 kg
Total weight at trailer axle = 22026.5 kg
Total weight at hauler rear axle = 15098.5 kg
Total weight at hauler front axle = 5971 kg

FEASIBLE SOLUTON: objective =30730.53
Packs of type 1 in zone 1 = 2
Packs of type 2 in zone 1 = 2
Packs of type 4 in zone 1 = 6
Packs of type 1 in zone 2 = 5
Packs of type 2 in zone 2 = 4
Packs of type 4 in zone 2 = 1
Packs of type 1 in zone 3 = 5
```

图 24.16 PFG 模型的 AMPL 输出结果

案例 8 Weyerhaeuser 木材切割及圆木分配的优化问题①

工具: 动态规划

应用领域: 木材厂工艺

问题描述

成材的大树被砍伐下来, 切割成圆木, 供不同的木材厂生产出不同的最终产品 (比如建筑木材、胶合板、圆片和纸张等). 供给每个木材厂的圆木规格 (例如长度、两端直径) 根据该木材厂生产的最终产品的要求有所不同. 用砍伐下来的 100 英尺高的树木, 按照木材厂要求的切割组合方式数可能非常大. 根据树木切割方式的不同, 造成的收益也不同. 本问题的目标是求出切割的组合方式, 使得总收入达到最大.

数学模型

模型的基本要求是, 我们不可能有一个最优的方案应用于某种"平均"的树木. 因为一般来讲, 砍伐下来的树其长度和直径都不同. 这就意味着我们必须根据每棵树来确定最优的切割方法和圆木段的分配.

这个模型的一个简化的假设是, 一棵砍伐下来的树的可用长度 L (英尺) 是某个最小长度 K(英尺) 的一个倍数. 此外, 从这棵树切下来的一段圆木的长度也是 K 的某个倍数. 这就是说, 圆木只能最小有 K 英尺, 最大长度为 NK 英尺长, 其中的 N 按照定义, $N \leqslant \frac{L}{K}$.

定义

$M=$ 需要圆木的木材厂数

$I=\dfrac{L}{K}$

$R_m(i,j)=$ 在木材厂 m, 从长度 iK 的一段树干的粗端切割下来的长度为 jK 的圆木所产生的收益, $m=1,2,\cdots,M$; $i=1,2,\cdots,I$; $j=1,2,\cdots,N$; $j\leqslant i$ $c=$ 在树木的 i 点进行切割的费用, $i=1,2,\cdots,I-1$

$$c_{ij}=\begin{cases} c, & 如果\ j<i \\ 0, & 如果\ j=i \end{cases}$$

从 c_{ij} 的定义可以看出, 若树干的长度 iK 等于所需要的圆木长度 jK, 则不需要切割.

为了理解符号 $R_m(i,j)$ 的含义, 图 24.17 给出了 $I=8$, $L=8K$ 的一棵树的示意图. 在 A 点和 B 点进行切割, 切出一段圆木供给木材厂 1, 两段给木材厂 2. 切割从树木的粗端开始, 在 A 点给木材厂 2 切割出圆木段 1, 这次切割对应于 $(i=8, j=3)$, 产生出收益 $R_2(8,3)$. 现在剩下的树干长度有 $5K$, 下一次在 B 点的

① 资料来源: M. R. Lembersky and U. H. Chi, "Decision Simulators Speed Implementation and Improve Operations," *Interfaces*, Vol.14, No.4, pp.1-15, 1984.

切割为木材厂 1 切出圆木段 2, 长度为 $2K$. 这段圆木对应于 $(i=5, j=2)$, 产生收益 $R_1(5,2)$. 剩下的长度为 $3K$ 的树干刚好等于给木材厂 2 的圆木段 3 的长度, 因此不需要再做切割. 相应的收益为 $R_2(3,3)$. 对应这个解的切割费用是 $c_{83}=c$, $c_{52}=c$, $c_{33}=0$.

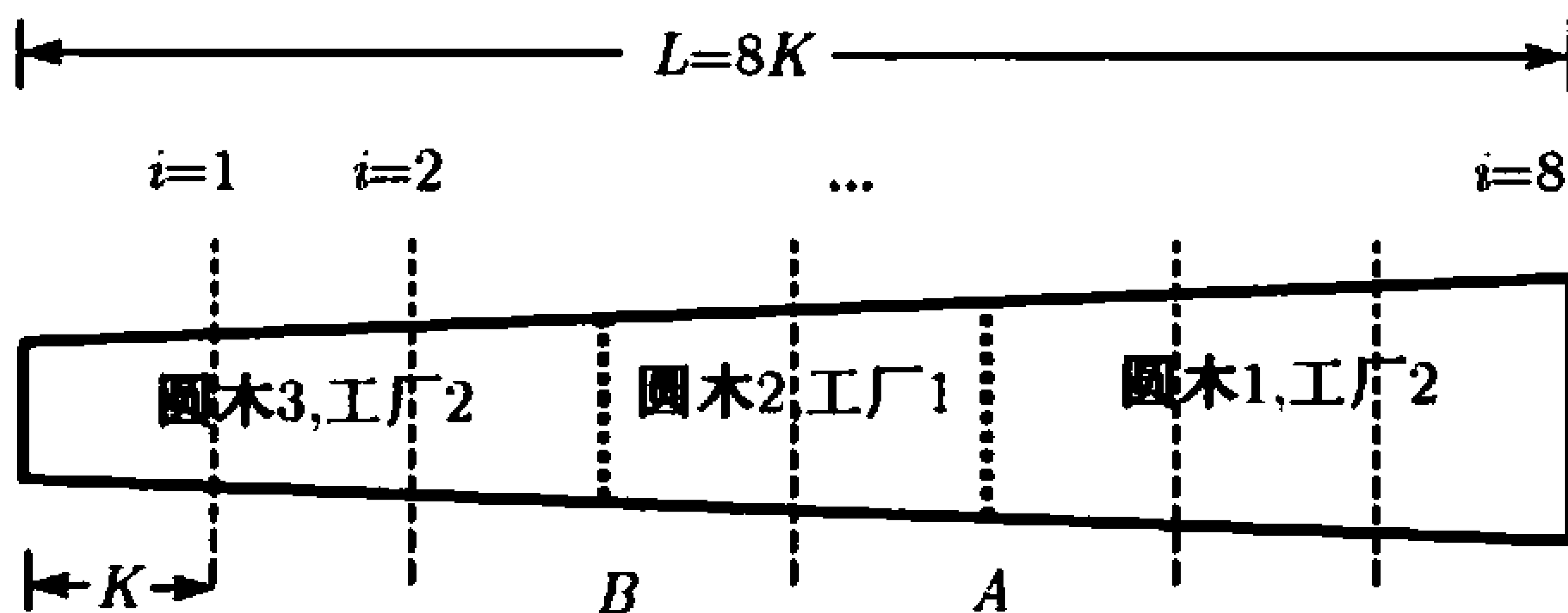


图 24.17 两个木材厂情形的典型解

这个问题可以用动态规划模型来刻画和求解.

令

$$f(i) = \text{当剩下的树干的长度为 } iK \text{ 时的最大收益}, \ i=1,2,\cdots,I,$$

则这个 DP 的递归方程可写成

$$f(0) \equiv 0$$

$$f(i) = \max_{\substack{j=1,2,\cdots,\min(i,N) \\ m=1,2,\cdots,M}} \{R_m(i,j) - c_{ij} + f(i-j)\}, \quad i=1,2,\cdots,I$$

上述递归方程的概念是, 给定树干的长度 iK, $f(i)$ 是一个收益函数, 等于切割一段长为 $j(\leqslant i)$ 圆木的收益, 减去切割费用, 加上从剩余长度为 $(i-j)K$ 的树干中得到的最佳累计收益.

实例计算

递归方程按照顺序 $f(1), f(2), \cdots, f(I)$ 来计算. 问题中有两个木材厂 $(M=2)$, 一棵树长度为 $L=12$ 英尺, 最短圆木为 $K=2$ 英尺长, 因此得出 $I=6$. 一次切割的费用是 $c=\$0.15$, 每个木材厂都接受长度为 2, 3, 4, 8, 10 英尺长的圆木, 这意味着 $N=5$. 图 24.18 是这个例子的电子表格程序解 (文件 excelCase8.xls). 基本的 DP 运算 (行 15~20) 有一部分是自动进行的, 当行 6~11 中的 $R_m(i,j)$ 改变了以后, 计算结果也自动修改. 所有的黑体数字需要手工输入. 这个电子表格程序局限于 $I=6$, $N=5$, $M=2$, 实际上只允许修改 $R_m(i,j)$ 的值①. 电子表格的 5~11 行给出了 $R_m(i,j)$, $j \leqslant i$ 的值. 注意对于一个指定的 $j=j^*$, $R_m(i,j^*)$ 的值随着 i 增加, 以反映圆木两端直径的增大.

① 这些电子表格公式应该提供足够的信息, 使它能适合其他的输入数据. 还可以用 VBA 的宏命令来开发一个一般性的电子表格程序, 用来指定矩阵 $R_m(i,j)$ 的大小, 并能自动进行所有的计算.

A	B	C	D	E	F	G	H	I	J	K	L	M	N
Input data													
Note: All italicized boldface data are supplied manually													
m=	2	K=	2	N=	5	L=	12	I=	6	c=	0.15		
Rm(i,j)		mill m=1						mill m=2					
j->>	1	2	3	4	5		1	2	3	4	5		
1	1					1	1.1						
2	1.1	1.15				2	1.1	2.3					
3	1.4	1.6	2.8			3	1.33	2.4	3.4				
4	1.9	1.9	3.9	4.1		4	2.1	3.3	4.2	4.1			
5	2.1	2.9	4.4	4.8	7.2	5	2.2	3.6	4.3	4.6	6		
6	2.1	3.5	4.7	6.1	8.3	6	2.2	4.5	4.4	5	6.3		
Calculations:												f(i)	(j*,m*)
i	j=1	j=2	j=3	j=4	j=5		j=1	j=2	j=3	j=4	j=5	$0.00	
1	1						1.1					$ 1.10	(1,2)
2	2.05	1.15					2.05	2.3				$ 2.30	(2,2)
3	3.55	2.55	2.8				3.48	3.35	3.4			$ 3.55	(1,1)
4	5.3	4.05	4.85	4.1			5.5	5.45	5.2	4.1		$ 5.50	(1,2)
5	7.45	6.3	6.55	5.8	7.2		7.55	7	6.5	5.55	6	$ 7.55	(1,2)
6	9.5	8.85	8.1	8.3	9.3		9.6	9.85	7.8	7.15	7.25		(2,2)
												Value=	$9.85
	m2		m1	m2	m2								
		i=1	i=2	i=3	i=4	i=5	i=6						

图 24.18 案例 8 的电子表格程序解

为了说明行 15~20 中的动态规划计算, 注意到每个阶段由一行组成, 因为在阶段 i, 该系统状态只有一个值, 即局部树干长度. 在阶段 $i = 1$, (剩余) 树干长度是 $1K$, 因此只能得到长度为 $1K$(即 $j = 1$) 的圆木. 还有, 因为没有做切割, $c_{11} = 0$. 因此有

$$
\begin{aligned}
f(1) &= \max\{R_1(1,1) - c_{11} + f(0), R_2(1,1) - c_{11} + f(0)\} \\
&= \max\{1 - 0 + 0, 1.1 - 0 + 0\} = 1.1
\end{aligned}
$$

阶段 $i = 1$ 时对应的最优决策是, 为木材厂 2 ($m^* = 2$) 切割一段长为 $1K(j^* = 1)$ 的圆木, 即 $(j^*, m^*) = (1, 2)$.

对阶段 2($i = 2$), 为这两个木材厂 ($m = 1$ 或 2) 提供的圆木长度都可以是 $1K$ 或 $2K$ (即, $j = 1$ 或 2), 因此,

$$
\begin{aligned}
f(2) &= \max\{R_1(2,1) - c_{21} + f(1), R_1(2,2) - c_{22} + f(0), R_2(2,1) - c_{21} + f(1), \\
&\qquad R_2(2,2) - c_{22} + f(0)\} \\
&= \max\{1.1 - 0.15 + 1.1, 1.15 - 0 + 0, 1.1 - 0.15 + 1.1, 2.3 - 0 + 0\} \\
&= \max\{2.05, 1.15, 2.05, 2.3\} = 2.3
\end{aligned}
$$

相应的最优决策是 $(j^*, m^*) = (2, 2)$, 要求为木材厂 2 切割一段长度为 $2K$ 的圆木.

类似地进行余下的计算, 如图 24.18 中的行 15~20 所示. 注意到, 电子表格中单元 B15:F20, H15:L20, M15:M20 是自动产生的. N15:N20 中的 (j^*, m^*) 在 15~20 行

自动计算完以后由手工输入. 行 15~20 手工输入的内容确定了 $f(i), i = 1, 2, \cdots, 6$.

从单元 N15:N20 读出下列最优解:

$$(i=6) \to (j^*, m^*) = (2,2) \to (i=4) \to (j^*, m^*) = (1,2) \to$$
$$(i=3) \to (j^*, m^*) = (1,1) \to (i=2) \to (j^*, m^*) = (2,2)$$

这个最优解要求, 在 $i = 2, 3, 4$ 点进行切割, 这棵树的总价值为 \$9.85.

实际应用

上述动态规划模型的结果可为实际操作工人在木材厂日常作业中使用. 这样, 就必须把模型编制成一个用户友好的系统, 其中那些复杂的动态规划计算对于用户来说是透明的. 这正是 Lemberskey and Chi [1] 在开发 VISION (Video Interactive Stem Inspection and OptimizatioN, 可视交互式树干检查与优化) 系统时的做法. 该系统带有一个伐木地区大量代表性树干样本的数据库. 这些数据包括树干的几何和质量数据 (例如, 枝节的位置), 以及不同长度和直径树干的价值 (美元). 另外还提供了不同木材厂所需要的质量特点.

VISION 系统的典型用户操作步骤如下.

第 1 步 操作工人可以从数据库里选择, 或利用 VISION 系统提供的图形功能创建一个样本树干, 从而在计算机屏幕上得到该树干的一个实际显示图. 木材厂所要求的圆木也可以从数据库中选择.

第 2 步 在屏幕上检查了树干后, 操作工人可以根据经验把该树干"切割"成若干圆木段. 接下来, 要求系统给出一个最优 DP 解. 在这两种情况下, 所得到的圆木段和相应的收益值都会以图形方式显示在屏幕上. 这样用户就有机会对这两种解进行比较. 尤其要仔细对 DP 给出的解进行检查, 保证所给出的圆木段符合质量要求. 若不符合的话, 用户可以对切割方案进行修正. 在每种情况下, 都会显示该树干相应的收益值用来进行比较.

在 VISION 系统中, DP 优化对于用户是完全透明的. 此外, 由于输出结果的交互式图形特点, 这个系统还可以用来对操作工人进行培训, 以提高操作者的决策技巧. 该系统的设计表明, 我们能够把复杂的数据模型嵌入到一个用户友好的计算机系统.

思考题

1. 假设在上面的例子中增加第 3 家木材厂, 它要求圆木段的长度有 4 英尺、6 英尺和 8 英尺. 这个新的木材厂有下面的 $R(i,j)$: $R(2,2) = \$2.50$, $R(3,2) = \$2.70$, $R(3,3) = \$3.10$, $R(4,2) = \$2.90$, $R(4,3) = \$3.30$, $R(4,4) = \$4.20$, $R(5,2) = \$3.10$, $R(5,3) = \$3.90$, $R(5,4) = \$4.60$, $R(6,2) = \$3.50$, $R(6,3) = \$4.20$,

$R(6,4) = \$5.10$. 修改电子表格程序 excelCase8.xls, 针对长度为 10′ 和 12′ 的树木, 求出最优解.

案例 9 计算机集成制造 (CIM) 设施的布局规划①

工具: AHP, GP

应用领域: 设施布局规划

问题描述

在某个学术机构, 要把一幢腾空的 5 100 平方英尺的大楼用来为工程学院建立一个 CIM 设施中心, 作为教学和科研实验室, 并建成一个产业技术进步中心. 学院向教师们征集了有关新建实验室布局规划的建议. 表 24.11 列出了中心内每个单位理想情况下以及最低要求情况下的面积. 走廊至少占 (实际) 总面积的 15%. 单元 1~5 主要用作教室, 每个教室必须能容纳 10~30 名学生.

表 24.11 中的理想面积比大楼总面积多出了约 1 300 平方英尺, 这就要求分配给某些单元的实际面积必须要减少. 面积的分配应考虑到这个新的 CIM 设施中心有 3 大功能: 教学、科研和产业服务. 教学需要在单元 1~5 的教室上课, 科研包括咨询项目和研究生课题, 这两项任务都要在实验室进行. 大家都认为, 该中心至少有 2 个咨询项目和 6 个研究生课题同时开展. 而产业服务主要是以继续教育研讨会的形式, 在接待大厅和报告厅进行 (单元 11).

表 24.11 理想的和最少的单元面积

单元	用途说明	理想面积 (ft^2)	绝对最小面积 (ft^2)
1	包装间	1 450	–
2	生产机床	148	–
3	计算机辅助设计/微机房	120	–
4	实物模拟实验室	530	–
5	智能控制工作站	130	–
6	原材料转化	165	–
7	组装	230	172
8	科研项目	400	–
9	技术员工作室	160	120
10	材料存放区	360	126
11	接待和报告厅	900	–
12	机器人系统	140	60
13	自动库存提取系统	500	200
14	传送系统	200	150
15	走廊空间	1 000	–
	总面积	6 433	

① 资料来源: C. Benjamin, I. Ehie, and Y. Omurtag, "Planning Facilities at the University of Missouri-Rolla," *Interfaces*, Vol.22, No.4, pp.95-105, 1992.

为了满足中心的教学 (T)、科研 (R) 和服务 (S)3 大功能, 学院成立了一个负责面积分配的委员会, 确定了该中心的以下 4 个主要目标.

目标 G1: 假设教室 1~5 每天的实验利用时间分别为 5,5,4,4,4 个小时, 这 5 个教室每天的学术利用率是 270 个学生实验-小时, 每次实验有 10~30 个学生参加.

目标 G2: 分配给实物模拟实验室 (单元 4) 的面积尽可能接近于要求的理想面积数 (= 530 ft^2).

目标 G3: 整个设施要能提供给共 15 个并行的研究项目和研究生课题, 随时至少有 2 个研究项目和 6 个研究生研究课题同时进行.

目标 G4: 继续教育报告厅 (单元 11) 至少要有 25 人以上的座位, 每位听众平均占 20 ft^2.

目标的优先级

这个问题涉及多个目标, 因此问题求解的第一步就是要确定这些目标的优先级. AHP(第 11.1 节) 对于这个目的而言是一种合适的分析工具. 规划委员会的 5 位成员每人做出了两组比较矩阵: 第一组对教学、科研和服务的主要面积进行排队, 而第二组矩阵对目标 G1 到 G4 进行打分. 表 24.12 和表 24.13 汇总了 5 位委员会成员所给出的打分结果.

表 24.12　对教学 (T)、科研 (R) 和服务 (S) 的比较矩阵

	委员 1			委员 2			委员 3			委员 4			委员 5		
	T	R	S	T	R	S	T	R	S	T	R	S	T	R	S
T	1	2	5	1	3	6	1	1	5	1	4	8	1	3	8
R	1/2	1	4	1/3	1	4	1	1	5	1/4	1	5	1/3	1	6
S	1/5	1/4	1	1/6	1/4	1	1/5	1/5	1	1/8	1/5	1	1/8	1/6	1

对这 5 位委员给出的比较矩阵, 利用几何平均化简为一个单一的比较矩阵, 因为几何平均保证了合并矩阵的元素 a_{ij} 和 a_{ji} 对所有的 i 和 j 满足所要求的倒数关系 $a_{ij} = \frac{1}{a_{ji}}$. 例如, 从表 24.12 中我们有

$$a_{12} = \sqrt[5]{2 \times 3 \times 1 \times 4 \times 3} = 2.352$$

$$a_{21} = \sqrt[5]{\frac{1}{2} \times \frac{1}{3} \times \frac{1}{1} \times \frac{1}{4} \times \frac{1}{3}} = 0.425$$

随后列出所得出的比较矩阵 (见文件 excelCase9a.xls):

$$\begin{array}{c} \\ \boldsymbol{A} = \begin{array}{c} \text{T} \\ \text{R} \\ \text{S} \end{array} \end{array} \begin{array}{c} \begin{array}{ccc} \text{T} & \text{R} & \text{S} \end{array} \\ \begin{pmatrix} 1.000 & 2.352 & 6.258 \\ 0.425 & 1.000 & 4.743 \\ 0.160 & 0.211 & 1.000 \end{pmatrix} \end{array}$$

$$
\boldsymbol{A}_{\mathrm{T}}=\begin{array}{c} \\ G1 \\ G2 \\ G3 \\ G4 \end{array}\begin{array}{c} \begin{array}{cccc} G1 & G2 & G3 & G4 \end{array} \\ \left(\begin{array}{cccc} 1 & 1.644 & 2.091 & 4.169 \\ 0.608 & 1 & 2.048 & 3.776 \\ 0.478 & 0.488 & 1 & 3.728 \\ 0.240 & 0.265 & 0.268 & 1 \end{array}\right) \end{array}
$$

$$
\boldsymbol{A}_{\mathrm{R}}=\begin{array}{c} \\ G1 \\ G2 \\ G3 \\ G4 \end{array}\begin{array}{c} \begin{array}{cccc} G1 & G2 & G3 & G4 \end{array} \\ \left(\begin{array}{cccc} 1 & 1.431 & 0.699 & 2.352 \\ 0.699 & 1 & 0.715 & 2.091 \\ 1.431 & 1.398 & 1 & 2.605 \\ 0.425 & 0.478 & 0.384 & 1 \end{array}\right) \end{array}
$$

$$
\boldsymbol{A}_{\mathrm{S}}=\begin{array}{c} \\ G1 \\ G2 \\ G3 \\ G4 \end{array}\begin{array}{c} \begin{array}{cccc} G1 & G2 & G3 & G4 \end{array} \\ \left(\begin{array}{cccc} 1 & 1.149 & 1.320 & 1.888 \\ 0.871 & 1 & 1.933 & 1.431 \\ 0.758 & 0.517 & 1 & 2.169 \\ 0.530 & 0.700 & 0.461 & 1 \end{array}\right) \end{array}
$$

表 24.13 目标 G1, G2, G3, G4 作为教学 (T)、科研 (R) 和服务 (S) 的函数的比较矩阵

教学(T)

	委员 1				委员 2				委员 3				委员 4				委员 5			
	G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4
G1	1	1	2	6	1	2	2	5	1	3	5	7	1	1	1	3	1	2	2	2
G2	1	1	3	8	1/2	1	3	6	1/3	1	2	4	1	1	2	4	1/2	1	1	1
G3	1/2	1/3	1	6	1/2	1/3	1	4	1/5	1/2	1	5	1	1/2	1	6	1/2	1	1	1
G4	1/6	1/8	1/6	1	1/5	1/6	1/4	1	1/7	1/4	1/5	1	1/3	1/4	1/6	1	1/2	1	1	1

科研 (R)

	委员 1				委员 2				委员 3				委员 4				委员 5			
	G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4
G1	1	3	2	4	1	1	2	1	1	2	1/2	3	1	1	1/4	2	1	1	1/3	3
G2	1/3	1	3	5	1	1	1	1	1/2	1	1/2	2	1	1	1/4	2	1	1	1/2	2
G3	1/2	1/3	1	2	1/2	1	1	1	2	2	1	4	4	4	1	3	3	2	1	5
G4	1/4	1/5	1/2	1	1	1	1	1	1/3	1/2	1/4	1	1/2	1/2	1/3	1	1/3	1/2	1/5	1

服务 (S)

	委员 1				委员 2				委员 3				委员 4				委员 5			
	G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4
G1	1	2	2	2	1	1	1	1	1	1	2	2	1	1	1	3	1	1	1	2
G2	1/2	1	3	3	1	1	1	1	1	1	3	2	1	1	1	3	1	1	1	1/2
G3	1/2	1/3	1	3	1	1	1	1	1/2	1/3	1	2	1	1	1	4	1	1	1	2
G4	1/2	1/3	1/3	1	1	1	1	1	1/2	1/2	1/2	1	1/3	1/3	1/4	1	1/2	2	1/2	1

对矩阵 $\boldsymbol{A}$, $\boldsymbol{A}_{\mathrm{T}}$, $\boldsymbol{A}_{\mathrm{R}}$, $\boldsymbol{A}_{\mathrm{S}}$ 运用 AHP 计算, 按照 excelCase9b.xls 中的目标 G1,

G2, G3, G4 排队, 结果为:

目标	权重值
G1	0.362
G2	0.285
G3	0.255
G4	0.098

排队结果表明, G1(教学) 得到最高优先级, G4(服务) 的优先级最低. 这意味着, 在求解相应的目标规划时, 优先级顺序是 $G1 \succ G2 \succ G3 \succ G4$.

电子表格程序中的 AHP 计算显示, 所有矩阵 $\boldsymbol{A}$, $\boldsymbol{A}_{\mathrm{T}}$, $\boldsymbol{A}_{\mathrm{R}}$, $\boldsymbol{A}_{\mathrm{S}}$ 的一致性比都在可接受的范围之内 (< 0.1).

目标规划的建模

令

$$a_i = \text{分配给单元 } i \text{ 的理想面积 (ft}^2\text{)}, \quad i = 1, 2, \cdots, 15$$

$$x_i = \text{分配给单元 } i \text{ 的实际面积与理想面积的比例}, \quad 0 \leqslant x_i \leqslant 1,$$
$$i = 1, 2, \cdots, 15$$

对 CIM 中心不同单元的实际分配面积涉及目标 G1, G2, G3, G4, 按照 AHP 计算得出了优先级次序, 即 $G1 \succ G2 \succ G3 \succ G4$. 除了灵活的目标约束以外, 每个目标还有一些严格的约束限制.

目标G1(每天的学生实验-小时目标数=270):

令

$$y_j = \text{参加实验 } j \text{ 的学生数}, \quad 10 \leqslant y_j \leqslant 30, \quad j = 1, 2, \cdots, 5$$

$$h_j = \text{实验室 } j \text{ 每天利用的小时数}, \quad j = 1, 2, \cdots, 5 (= 5, 5, 4, 4, 4\text{个小时})$$

$$c_1 = \text{每个学生在实验室单元}1, 2, \cdots, 5\text{所需要的面积数}(= 10 \text{ ft}^2)$$

$$H = \text{所有实验室每天使用小时的目标数}(= 270)$$

$$p_1 = \text{目标 G1 没有实现的量}$$

$$q_1 = \text{目标 G1 超额实现的量}$$

因此, G1 的问题可表示为

$$\min \ G1 = p_1$$
$$\text{s.t.} \quad \left.\begin{array}{l} y_j \leqslant \dfrac{a_j x_j}{c_1} \\ 0 \leqslant x_j \leqslant 1 \\ 10 \leqslant y_j \leqslant 30 \end{array}\right\}, \quad j = 1, 2, \cdots, 5$$

$$\sum_{j=1}^{5} h_j y_j + p_1 - q_1 = H$$
$$p_1,\ q_1 \geqslant 0,\ \text{所有的 } y_j \text{ 为整数}$$

目标G2(分配给实物模拟实验室的理想面积, $x_4 = 1$):

G2 的问题如下:

$$\begin{aligned} \min\ & G2 = p_2 \\ \text{s.t.}\ & x_4 + p_2 - q_2 = 1 \\ & 0 \leqslant x_4 \leqslant 1 \\ & p_2,\ q_2 \geqslant 0 \end{aligned}$$

目标G3(并行研究项目和课题的总数 $\geqslant 15$):

令

w_1 = 研究项目数($\geqslant 2$)
w_2 = 研究课题数($\geqslant 6$)
b_1 = 每个研究项目所需要的面积($= 36\ \text{ft}^2$)
b_2 = 每项研究课题所需要的面积($= 48\ \text{ft}^2$)

G3 的问题为

$$\begin{aligned} \min\ & G3 = p_3 \\ \text{s.t.}\ & w_1 + w_2 + p_3 - q_3 = 15 \\ & b_1 w_1 + b_2 w_2 \leqslant a_8 x_8 \\ & w_1 \geqslant 2 \\ & w_2 \geqslant 6 \\ & 0 \leqslant x_8 \leqslant 1 \\ & p_3,\ q_3 \geqslant 0 \end{aligned}$$

目标G4(继续教育教室容量 $\geqslant 25$ 人):

令

z = 每次继续教育研讨班的参加人数
d = 每位听课人员需要的平均教室面积($= 20\ \text{ft}^2$)

G4 的问题为

$$\min\ G4 = p_4$$

$$\begin{aligned}\text{s.t.}\quad & z + p_4 - q_4 = 25\\ & dz \leqslant a_{11}x_{11}\\ & 0 \leqslant x_{11} \leqslant 1\\ & z \geqslant 0,\ \text{取整数}\\ & p_4,\ q_4 \geqslant 0\end{aligned}$$

适用于所有目标的强制性约束是

(1) 可分配的总面积：

$$\sum_{i=1}^{15} a_i x_i = A$$

(2) 需要的最小面积：

$$x_7 \geqslant \frac{172}{230},\quad x_9 \geqslant \frac{120}{160},\quad x_{10} \geqslant \frac{126}{360},\quad x_{12} \geqslant \frac{60}{140},\quad x_{13} \geqslant \frac{200}{500},\quad x_{14} \geqslant \frac{150}{200}$$

(3) 走廊要求：

$$a_{15}x_{15} \geqslant 0.15A$$

模型求解

可以通过把所有目标合并成一个目标函数, 即用 AHP 对每个目标 G1, G2, G3, G4 的优先级赋以权重 0.337, 0.288, 0.240, 0.134 来求解这个问题, 即,

$$\min\ G = 0.362p_1 + 0.285p_2 + 0.255p_3 + 0.098p_4$$

这个合并问题的约束, 包括所有 G1 到 G4 的约束条件, 加上共有的强制性约束.

文件 amplCase9.txt 提供了该问题的一个 AMPL 程序模型, 图 24.19 给出了解的结果. 该结果表明, 目标 G1 被每天 3 个学生–小时数实现得过多, 目标 G3 有 6 个课题未能达到目标, 而目标 G4 刚好实现.

灵敏度分析

用于 AHP 计算的提供了原始数据的偏好矩阵, 是参与打分过程的每个人的一种主观评价. 可以用灵敏度分析来研究数据偏差对解的影响. 有意思的是, 用 AMPL 模型所作的实验显示, 不管我们对目标所赋予的权重如何, 所得到的解仍保持不变. 的确, 应用于目标权重的 AMPL 的灵敏度分析提供了如下结果:

	p. 下降	p. 不变	p. 上升
G1	0	0.362	1e+20
G2	0	0.285	1e+20
G3	0	0.255	1e+20
G4	0	0.098	1e+20

这些结果说明, 对任何 0 到无穷大之间的权重值, 最优解保持不变. 这一现象表明有一种特殊的可行解空间, 可能只有一个点构成, 因此最优解与目标函数无关.

```
Projects:
Contract   3
Research   6

Space allocation:
Unit     Factor        Actual sq ft
 1       0.094           135
 2       1.000           148
 3       0.833            99
 4       1.000           530
 5       1.000           130
 6       1.000           165
 7       1.000           230
 8       1.000           400
 9       1.000           160
10       1.000           360
11       1.000           900
12       1.000           141
13       1.000           500
14       1.000           200
15       1.000          1000
            Total=      5099

Student enrollments in labs:
Lab 1     11
Lab 2     14
Lab 3     10
Lab 4     17
Lab 5     10

Deviational variables:
      p       q
G1    0       3
G2    0       0
G3    6       0
G4    0       0
```

图 24.19 面积分配问题的 AMPL 输出结果

从上面的结果中, 我们得出的主要经验就是, 假如我们能用我们对目标权重所做出的最好的“推测”来构造出这个目标规划问题, 然后对这些权重的改变可能对解的影响进行仔细的研究的话, 那么前面进行的所有 AHP 计算都可以省去了. 当然, 我们建立模型的方法是合理的, 而且我们后来才发现这一现象. 但不管怎样, 这个经验说明, 在不影响最终建议方案准确性的前提下, 如果能够找到更简单的方法, 我们就不应该采用较复杂的分析工具 (例如 AHP).

思考题

1. 假设大楼的实际面积有 5 600 ft^2, 且进一步假定单元 11(报告厅) 要求的理想面积为 1 000 ft^2, 求最优解.

案例 10 旅店客房的预定上限问题①

工具: 决策树分析

应用领域: 旅店业

问题描述

La Posada 旅店共有 300 间客房, 旅店的顾客既有商务人士, 也有休闲旅游者. 客房可以事先预定 (通常给休闲游客), 并享受优惠. 而商务出差人士总是订房比较晚, 付全价. 因此, La Posada 旅店必须对租给休闲游客的折扣价客房设定一个*预定上限*, 以便从付全价的商务顾客那里获得更大的收益.

数学模型

令 N 表示所有客房数量, 并假定按全价销售的客房保留数为 $Q+1, 0 \leqslant Q < N$. 相应的预定上限 (按折扣价销售的客房数) 就是 $N-Q-1$. 图 24.20 是该问题的示意图.

为了确定是否应将全价房预留数从 $Q+1$ 减少为 Q, 我们利用图 24.21 中的决策树. 设 D 是一个随机变量, 表示过去的或者预测的对全价 (商务人士) 房的需求量. 进一步, 令 c 为全价价格, d 是折扣价格 ($d < c$). 把保留数从 $Q+1$ 减少到 Q 的决策意味着房间 $Q+1$ 肯定能够按照折扣价 d 销售出去, 因为这种机会很大. 另一方面, 不降低全价房预留数预会产生两种不确定的结果: 假如对商务房的需求量大于或等于 $Q+1$, 则房间 $Q+1$ 就会按照全价 c 卖出; 否则的话, 这间客房根本就出租不出去. 相应的概率分别为 $\mathrm{P}\{D \geqslant Q+1\}$ 和 $\mathrm{P}\{D \leqslant Q\}$. 这样就可以得到下面的结论: 若

$$d \geqslant c\mathrm{P}\{D \geqslant Q+1\} + 0\mathrm{P}\{D \leqslant Q\}$$

① 资料来源: S. Netessine and R. Shumsky, “Introduction to the Theory and Practice of Yield Management,” *INFORMS Transactions on Education*, Vol.3, No.1, pp.20-28, 2002.

或

$$\mathrm{P}\{D \leqslant Q\} \geqslant \frac{c-d}{c}$$

则应该采取把预留数降低到 Q 的决策.

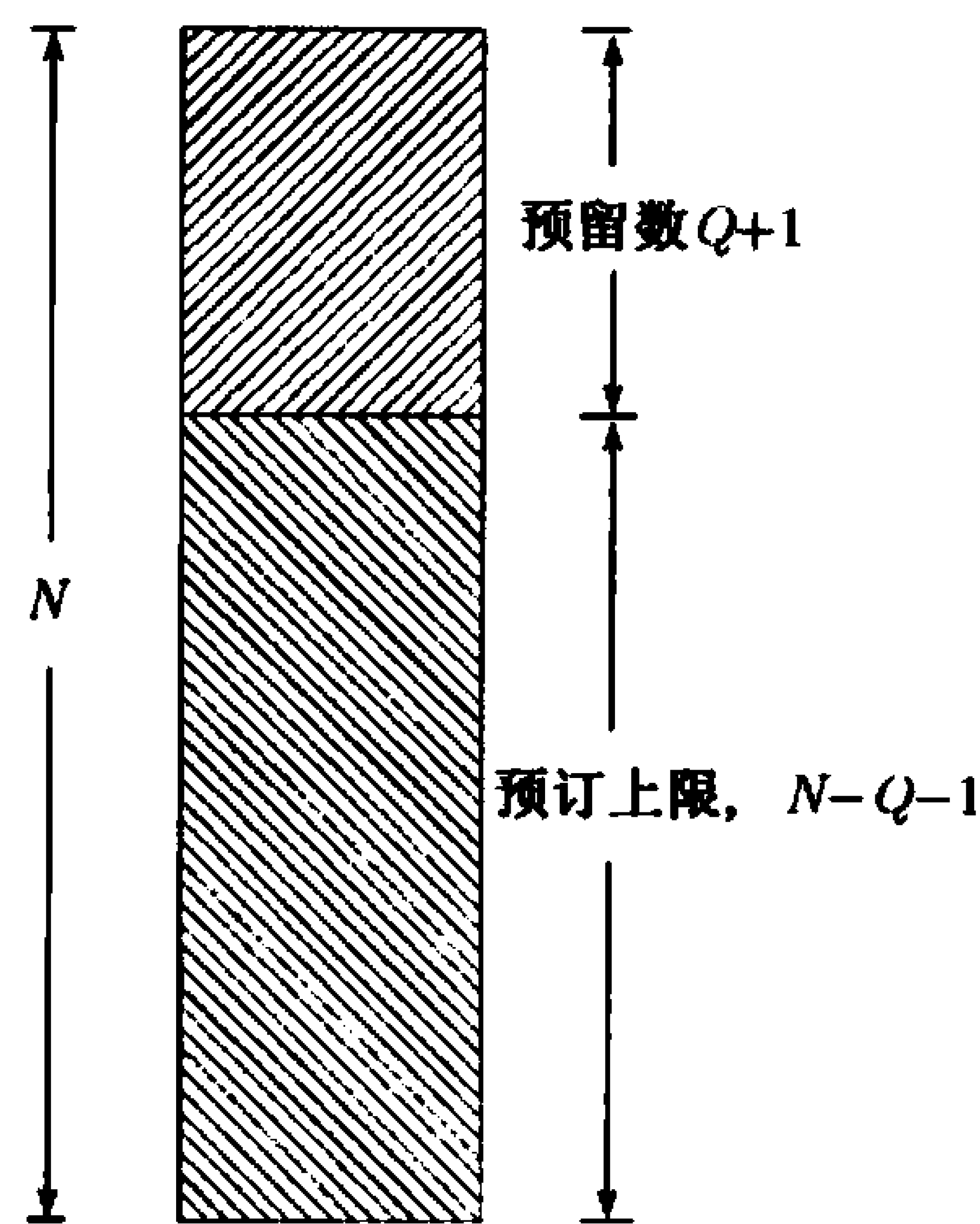


图 24.20 预订上限和预留数

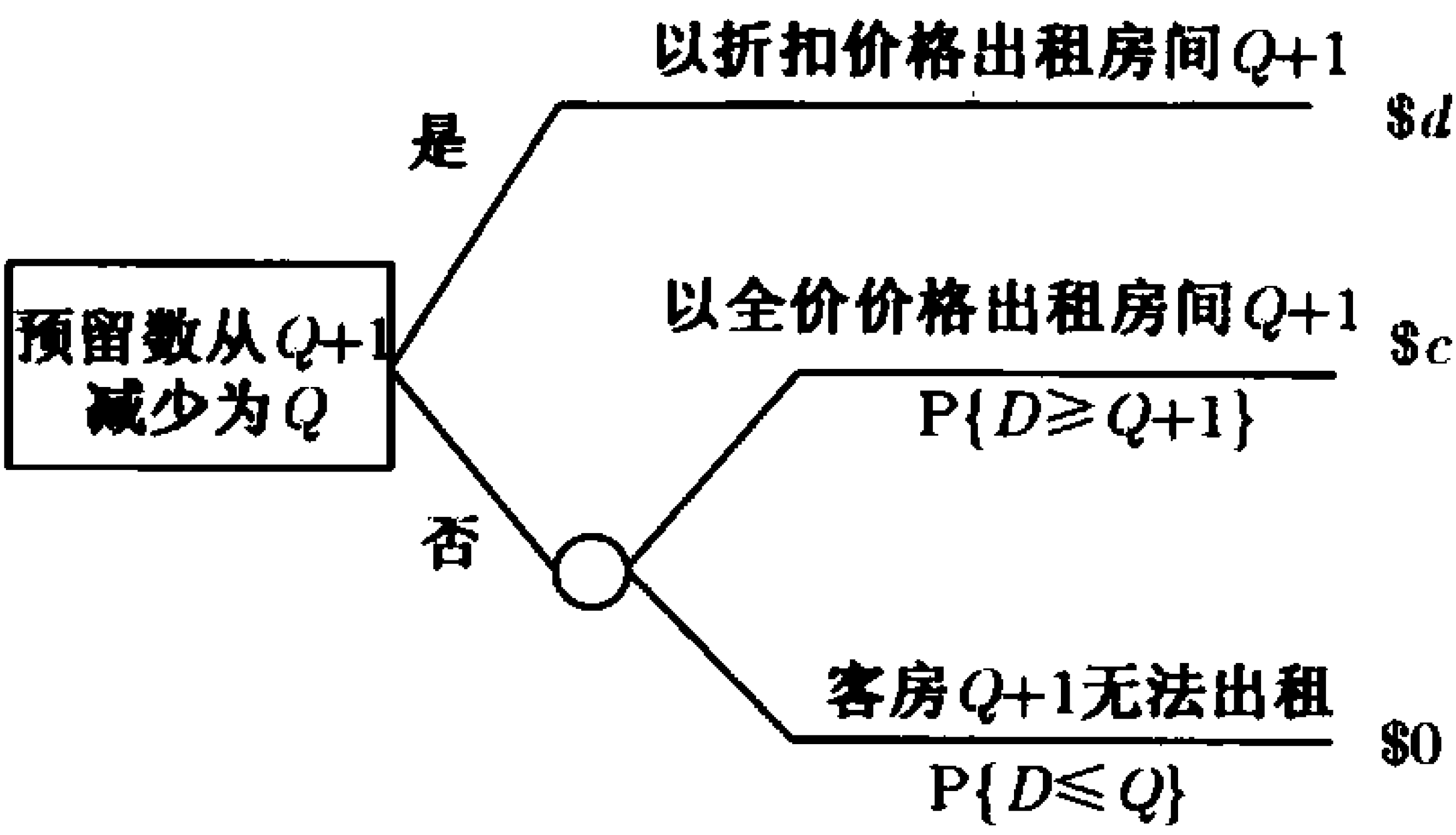


图 24.21 求预留客房数 Q 的决策树

如果我们知道了需求 D 的分布, 以及单位价格 c 和 d, 就可以决定预留数 Q 了.

数据采集

决定预留房间数所需要的最重要的信息就是对全价客房的需求分布. 为此可以采用某个时间段上的历史数据. 使用一批客房 Q 按全价预定的天数就能估计出需求概率 $\mathrm{P}\{D = Q\}$, 以此可以确定出累积概率来. 表 24.14 给出了求需求分布的数据. 前面两列是原始数据.

我们通过下面的例子来说明如何使用表 24.14 中的信息. 假设全价是 \$159, 折扣价为 \$105, 根据

$$\mathrm{P}\{D \leqslant Q\} \geqslant \frac{159-105}{159} = 0.339\,62$$

来决定预留上限. 表 24.14 中的累积概率列表明预留数应该为 $Q = 79$ 间客房.

结论

本问题的思路可以类似地扩展到设定飞机票的预售上限. 此外, 除了采用预定上限, 我们还可以对这种分析方法进行修改, 用来设定多个水平的预定上限, 让折扣价格随着预定日期的临近而上涨. 用于这个模型的最重要的信息就是对需求数据的可靠估计.

表 24.14　P{D=Q} 和 P{$D \leqslant Q$} 的计算

客房数 Q	需求天数	P{$D=Q$}	P{$D \leqslant Q$}
0 − 70	12	0.0975 6	0.097 561
71	3	0.024 39	0.121 95
72	3	0.024 39	0.146 34
73	2	0.0162 6	0.162 60
74	0	0.000 00	0.162 60
75	4	0.032 52	0.195 12
76	4	0.032 52	0.227 64
77	5	0.040 65	0.268 29
78	2	0.016 26	0.284 55
79	7	0.056 91	0.341 46
80	4	0.032 52	0.373 98
81	10	0.081 30	0.455 28
82	13	0.105 69	0.560 98
83	12	0.097 56	0.658 54
84	4	0.032 52	0.691 06
85	9	0.073 17	0.764 23
86	10	0.081 30	0.845 53
> 86	19	0.154 47	1.000 00

思考题

1. 假设没有卖出的商务房会有一个 50% 的机会由休闲游客按照高价购买，这对于确定预定上限有什么影响？
2. 给出该模型所基于的基本假定，这些假定符合实际吗？
3. 假定对商务旅客采用两种预留数水平：按照每间房间 \$$c_1$ 预留 Q_1 间房，按每间房间 \$$c_2$ 预留 Q_2 间房，$c_1 > c_2$. 余下的房间按照每间房 \$$d_1$ 的折扣价销售. 建立一个模型，求 Q_1 和 Q_2，并说出所有的假设前提.

案例 11　Casey 问题：对一次全新化验结果的解释和评估①

工具：决策树 (贝叶斯概率)

应用领域：医疗化验

问题描述

有一个新生儿名叫 Casey，对她的一次初筛检查中查出她患有一种 C14:1 酶缺乏症. 这种酶用于消化一种特殊形式的长链脂肪，一旦缺乏这种酶，人就会产生严重的疾病，也可能导致原因不明的死亡 (一般称为婴儿猝死综合症，简称 SIDS). 以

① 资料来源：J. E. Smith and R. L .Winkler, "Casey's Problem: Interpreting and Evaluating a New Test," *Interfaces*, Vol.29, No.3, pp.63-76, 1999.

前对大约 13 000 个新生儿做过这项化验, Casey 是第一例呈阳性化验结果的新生儿. 对于这种疾病尚没有已知的治疗方法, 但她的医生称, 有可能通过限制饮食来控制住病情的发展. 虽然只凭这一项化验还不能得出确定的诊断, 但是这种情况的罕见性还是让她的医生相信, Casey 有 80%~90% 的机会患上了这种缺乏症.

由于医生评估的高概率 (0.8~0.9) 只是基于初筛化验的结果, 就需要进行一套系统化的分析, 来解释和评估医生关于这种病情所给出的定量判断. 尤其是, 在已知 Casey 被化验出阳性以后, 我们要知道她确实患上这种 C14:1 缺乏症的概率是多少?

由于对这种病症缺乏足够的数据, 使得决策的过程变得复杂. 仅有的确切信息是, Casey 的阳性结果在过去的 13 000 次化验中还是首例. 医疗纪录也说明了, 在发生婴儿猝死综合症的病例中, 尸体解剖都表明这些死亡不是 C14:1 缺乏症造成的. 在这方面, 这种缺乏症化验没有 (真或假) 阳性结果的历史资料. 在这种情况下的任何决策必须只能依赖于对所需要数据的“合理的”估计.

分析

医生估计, 总体上来讲, 新生儿 C14:1 缺乏症的发病率大约是 $\frac{1}{40\,000}$. 但是, Casey 的情况太少见了, 她这种情况的*发病率*有可能降低到 250 000 人中才有一例. 其他能够估计的数据都是假结果, 其中新生儿被*错误地*检验出阳性或阴性. 从假阳性结果开始, 其他已有历史数据的初筛化验, 例如 HIV、毒品或者氨基胎蛋白化验等, 有可能用来作为指导. 这些化验的假阳性率从 $\frac{1}{100}$ 到 $\frac{1}{1\,000}$ 不等. 然而对于 Casey 的化验, 这个范围似乎太大了, 因为历史数据显示 13 000 中才有一例. 因此我们有理由推断, 这个化验可能有 $\frac{1}{20\,000}$ 的假阳性率. 对于当新生儿真的患有这种缺乏症时出现的假阴性率, 据估计不会高, 因为这种化验都是按照严格的标准来管理的, 据此我们采用 $\frac{1}{1\,000}$ 作为出现这种情况的一个接近实际的“推测”.

可以用贝叶斯定理来分析上述决策问题, 目的是在已知 Casey 化验的结果呈阳性的情况下, 计算出她患有 C14:1 缺乏症的后验概率. 图 24.22 总结了利用上述估计值所得出的相应问题的概率.

可以用下面的贝叶斯公式, 算出已知 Casey 化验呈阳性情况下她患有 C14:1 缺乏症的后验概率:

$$\begin{aligned}P\{C14:1|+ve\text{ 化验}\} &= \frac{P\{+ve|C14:1\}P\{C14:1\}}{P\{+ve|\ C14:1\}P\{C14:1\}+P\{+ve|\text{noC14}:1\}P\{\text{noC14}:1\}}\\ &= \frac{0.999\times 0.000\,004}{0.999\times 0.000\,004+0.000\,005\times 0.999\,996}=0.074\end{aligned}$$

这一计算结果说明, 已知在 Casey 化验的结果呈阳性的情况下, 她患有 C14:1 缺乏症的概率只有 0.074. 虽然这个概率比医生提出的 0.8 ~ 0.9 要小得多, 但足以引起关注, 因为这个数意味着发生率为 $\frac{1}{13.5}$. 但是, 医生的估计怎么会错这么多呢?

有没有可能图 24.22 中的发病、假阴性、假阳性概率估计得极不准确呢? 因为这些概率都不是根据可靠的历史数据算出来的, 因此这似乎是有问题的.

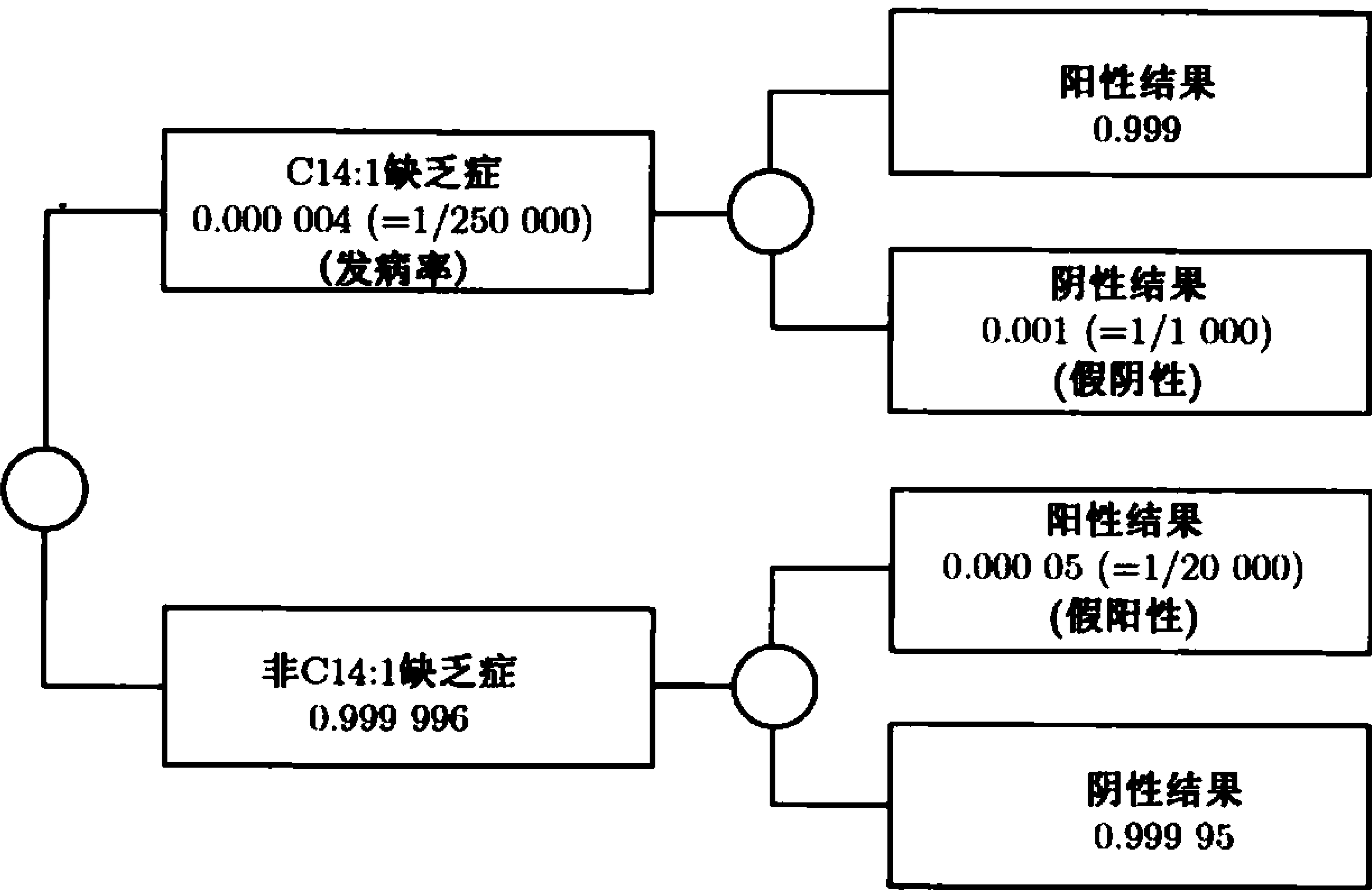


图 24.22　Casey 的化验概率的决策树

检验这些给出的结果的一种方法是, 进行灵敏度分析计算, 来研究改变发病率、假阴性率和假阳性率对所要求的后验概率会产生什么影响. 表 24.15 列出了作灵敏度分析的范围.

表 24.15　灵敏度分析的范围

	发生率		
	最大值	基准值	最小值
发病率	1/100 000	1/250 000	1/1 000 000
假阴性	1/100	1/1 000	1/1 000 000
假阳性	1/5 000	1/20 000	1/1 000 000

我们通过改变一项比率的范围, 同时把其他两个率保持在基准水平来算出后验概率. 利用贝叶斯公式 (文件 excelBayes.xls), 得出表 24.16 的计算结果.

表 24.16　灵敏度分析的范围对应的后验概率

	对应的后验概率			灵敏度系数
	最大发生率	基准发生率	最小发生率	
发病率	0.166 5	0.074 0	0.019 6	2.25
假阴性	0.073 4	0.074 0	0.074 0	1.00
假阳性	0.019 6	0.074 0	0.800 0	10.81

表 24.16 最后一列的灵敏度系数度量了后验概率对发病率、假阴性率和假阳

性率改变后的灵敏程度. 它是最大后验概率与相对于基准率的后验概率之比. 例如, 对于假阳性情况, 这个比例是 0.8/0.074 0 = 10.81. 我们主要关心的是造成后验概率增大的假阳性率值, 即发生假阳性和假阴性的最小发生率, 以及发病情况下的最大发生率.

这个灵敏度系数说明, 假阳性率对后验概率的影响最大, 而其他两个发生率不那么重要. 这对我们的警示是, 按照 $\frac{1}{1\ 000\ 000}$ 的假阳性率, 患 C14:1 缺乏症的后验概率就能达到医生所估计的高值 (= 0.8 ~ 0.9). 这就意味着, 对于假阳性发生率, 还需要做进一步的分析.

在前面的分析中, 我们根据原来的 13 000 次化验没有发生假阳性这一现象, 估计出基准假阳性发生率为 $\frac{1}{20\ 000}$. 另一种方法是, 假定我们没有这 13 000 例初筛化验的任何先验知识的前提下, 我们要对这一发生率做出一个先验的"推测". 首先假设, 在 1 000 个化验样本中发生了一例假阳性, 这等价于说, *期望的*假阳性发生率是 0.001. 现在, 根据我们知道的信息, 在 13 000 例实际化验中没有出现假阳性, 我们就可以把相应的假阳性发生率修改成 14 000 中出现一例 (= 1 000 + 13 000). 因此相应的 C14:1 缺乏症的后验概率就可以计算为

$$\begin{aligned} P\{C14:1\text{缺乏症}|+ve\ \text{化验}\} &= \frac{0.999 \times 0.000\ 004}{0.999 \times 0.000\ 004 + (1/14\ 000) \times 0.999\ 996} \\ &= 0.052\ 98 \approx 0.053 \end{aligned}$$

假如我们把样本数量从 1 000 减小到 100, 同时让 0.001(1 000 次发生 1 例) 的*期望*假阳性发生率保持不变, 则相应的假阳性发生率就变成了 13 100 中发生 0.1 例, 而且相应的后验概率为

$$\begin{aligned} P\{C14:1\text{缺乏症}|+ve\ \text{化验}\} &= \frac{0.999 \times 0.000\ 004}{0.999 \times 0.000\ 004 + (0.1/13\ 100) \times 0.999\ 996} \\ &= 0.343\ 6 \end{aligned}$$

表 24.17 比较了不同期望假阳性发生率的后验概率值.

表 24.17 3 个假阳性发生率的后验概率比较

期望的假阳性发生率	对应样本规模 n 的后验概率		
	$n = 100$	$n = 1\ 000$	$n = 10\ 000$
0.01	0.049 7	0.005 56	0.000 92
0.001	0.343 6	0.052 98	0.009 11
0.000 001	0.998 1	0.982 40	0.478 90

这些计算给出了一个有意思的结果. 对于达到医生提出的 0.8 ~ 0.9 的后验概率, 必出现小样本量 ($n \leqslant 1\ 000$) 与非常低的期望假阳性发生率的组合方案, 这是一个极不可能的组合. 这里这个结论是, 医生对 Casey 患有 C14:1 缺乏症的结论是夸大了.

上述分析并不能说医生错了, 也不说明 C14:1 缺乏症并不存在. 它只是指出一个事实, 即最初的化验结果不能用于对病情做出一个明确的诊断. 解决这个问题的唯一方法只能是进行进一步的检查, 实际上对于 Casey 的情形就是这样做的.

思考题

1. 用 excelBayes.xls 检验表 24.17 中的后验概率值.

案例 12 莱德杯决赛中高尔夫球手的出场顺序安排[①]

工具: 对策论

应用领域: 体育比赛

问题描述

一次高尔夫球循环赛的最后一天决赛中, 两支队争夺冠军. 每个队的队长必须提交一份高尔夫球手的出场顺序单 (*出场名单*), 这个名单也就确定了每一局的比赛对手. 因此, 假如 A 队和 B 队的出场名单为 $(A1, A2, \cdots, An)$ 和 $(B1, B2, \cdots, Bn)$, 则各局的比赛对手就是 A1 对 B1, A2 对 B2, $\cdots$, An 对 Bn. 我们似乎可以假定, 假如两个水平相当的球手在他们的分别出场名单中排在同样的位置, 那么每个球手赢得比赛的机会为 50:50. 当排在前面的球手和排在后面的球手比赛时, 获胜的概率就会增加. 为了说明这个概念, 表 24.18 给出了 A 队的一名高尔夫球手获胜的一组概率.

表 24.18 A 队获胜的概率

	B1	B2	B3
A1	0.50	0.60	0.70
A2	0.40	0.50	0.60
A3	0.30	0.40	0.50

我们的目标是建立一套系统化的方法, 看看所采用的比赛顺序的思路是对还是不对. 我们假定, 每局比赛不能平局, 而且所有的队员都能顶住压力, 也就是说, 一名队员的表现在整个比赛中始终保持不变.

分析

可以用二人常和对策方法来分析这个问题. 一个出场名单代表每个队员的一个可能的策略, 用获胜概率来表示一个参赛队的收益. 对于一场有 3 个队员参加的比

① **资料来源:** W. Hurley, "How Should Team Captains Order Golfers on the Final Day of the Ryder Cup Matches?" *Interfaces*, Vol.32, No.2, pp.74-77, 2002.

赛, 每个队有 6(3!) 种可能的策略 (出场名单), 如表 24.19 所示.

表 24.19 A 队和 B 队的出场顺序

策略	出场顺序	
	A 队	B 队
1	A1,A2,A3	B1,B2,B3
2	A1,A3,A2	B1,B3,B2
3	A2,A1,A3	B2,B1,B3
4	A2,A3,A1	B2,B3,B1
5	A3,A1,A2	B3,B1,B2
6	A3,A2,A1	B3,B2,B1

为了说明如何求 A 队的收益矩阵 (概率), 考虑 A 队采用策略 1, B 队采用策略 5. 按照各自的策略, 让 A1 与 B3 比赛, 让 A2 对 B1, 让 A3 对 B2, 比赛采用三局两胜制. A 队获胜的概率计算为

$$0.7\times0.4\times0.4+(1-0.7)\times0.4\times0.4+0.7\times(1-0.4)\times0.4+0.7\times0.4\times(1-0.4)=0.496$$

整个收益矩阵 (概率) 用 excelCase12.xls 来计算, 如表 24.20 所示.

表 24.20 A 队的收益矩阵

	1	2	3	4	5	6
1	0.500	0.500	0.500	0.504	0.496	0.500
2	0.500	0.500	0.504	0.500	0.500	0.496
3	0.500	0.496	0.500	0.500	0.500	0.504
4	0.496	0.500	0.500	0.500	0.504	0.500
5	0.504	0.500	0.500	0.496	0.500	0.500
6	0.500	0.504	0.496	0.500	0.500	0.500

这个对策的解 (用 toraCase12.txt) 要求, 每个队要把选定的对策进行混合 (该问题还有其他解). 对策值为 0.5, 对 A 队有利.

思考题

1. A 队和 B 队该如何排列它们的策略?
2. 若 A 队获胜的概率为

$$\begin{pmatrix} 0.4 & 0.2 & 0.7 \\ 0.7 & 0.5 & 0.8 \\ 0.8 & 0.9 & 0.6 \end{pmatrix}$$

求出两名队员的最优策略.

案例 13 戴尔供应链的库存决策[①]

工具: 库存模型

应用领域: 库存控制

问题描述

戴尔公司实行一种直销式的经营模式, 按照这种模式它向顾客直接销售个人电脑. 当收到来自美国客户的一份订单后, 公司将订单配置要求发送到设在得克萨斯州奥斯汀的某个制造厂, 在那里 8 个小时之内就能对订单要求的产品进行组装、测试和包装. 戴尔公司自己的库存很小, 因为它要求东南亚地区的供应商保持一种所谓的“周转库存”, 设在制造厂附近的*周转仓库*(零部件仓库) 那里. 这些周转仓库属于戴尔公司, 租给供应商使用, 然后, 戴尔从周转仓库那里“提出”所需要的零部件, 而供应商的责任是及时补充库存, 以满足戴尔的预计需求. 短缺的零件通过较昂贵的连夜运送得以补给. 虽然戴尔并不拥有周转仓库中的货物, 但库存的费用通过零部件的价格间接传递给顾客. 因此, 库存的降低就可以通过降低产品成本直接让戴尔的顾客受益. 本问题的目标就是提出合理的周转库存量, 使得库存成本达到最小.

问题分析

问题分析始于选择数学模型, 然后收集数据, 用于计算模型, 最后对得到的结果进行分析, 做出最终的决策.

我们的研究集中在 XDX 上, 它是一种主要的个人电脑零部件. 该部件由多家供应商提供, 按批量送货给周转库存, 以减少固定的订货费. 每家供应商的库存策略是, 一旦库存水平下降至某个水平, 就送一批新货. 但是, 公司注意到, 供应商之间的批量大小和再订货水平差别很大, 所以首先需要让这些供应商的订货策略一致起来. 为了表达这个情形, 所选择的模型是 (Q, R) 策略, 其中 Q 是批量大小, R 为再订货水平. 历史数据显示, 订货量 Q 不变, 订货策略的不一致主要是由于再订货水平, 以及从发出订单到到货之间的提前时间有所不同. 因此, 所作的一个决策是, 把 Q 保持在现在的水平上不变, 而分析主要集中在如何确定 R 上.

可以合理地假定, 提前时间期间的需求量 x 服从均值为 μ、标准差为 σ 的正态分布, 因此提前时间需求的标准正态分布定义为

$$z = \frac{x - \mu}{\sigma}$$

① **资料来源**: I. Millet, D. Parente, J. Fizel, and R. Venkataraman, “Inventory Decisions in Dell's Supply Chain,” *Interfaces*, Vol.34, No.3, pp.191-205, 2004.

接下来, 令 α 表示提前时间期间戴尔可以接受的用完库存的概率, 定义 Z 使得 $P\{z \leqslant Z\} = 1 - \alpha$, 则再订货水平 R 就是

$$R = \mu + Z\sigma$$

可以用一个近似的单周期报亭库存模型来启发式地求出 Z(见 15.2 节), 这种单周期模型的合理性依据是, 零部件 XDX 具有一定的有效期, 过期报废. 已知 p 是对每件未满足需求的惩罚费用, h 为每单位过剩库存的储存费用, 我们有

$$\Phi(Z) \equiv P\{z \leqslant Z\} = \frac{p}{p+h}$$

函数 $\Phi(Z)$ 是标准正态分布的 CDF. 因此有

$$Z = \Phi^{-1}\left(\frac{p}{p+h}\right)$$

提前时间期间需求量的平均值 μ 和标准差 σ 是两个随机变量的函数: 提前时间和每单位时间需求量. 在某种独立假设的前提下, 可以证明[①]

$$\mu = DT, \qquad \sigma = \sqrt{D^2\sigma_t^2 + T\sigma_d^2}$$

其中

D = 单位时间平均需求件数, σ_d = 需求量标准差

T = 平均提前时间量, σ_t = 提前时间的标准差

基于这个模型, 若固定订货量为 Q 件, 该库存策略是"一旦当周转库存水平下降到 R 件时, 就订 Q 件 XDX".

数据采集

数据的收集来自 A, B, C 3 家供应商, 他们提供不同的但是完全可以互换的 XDX 零部件, 来补充戴尔的周转库存. 模型所需要的数据采自 1998 年 12 月 1 日到 1999 年 5 月 27 日之间, 分两组: (1) 与所估计的储存费用 h 和惩罚费用 p 有关的数据; (2) 用来估算提前时间期间需求平均值以及方差的数据.

储存费用 h 是根据两次连续供货之间的时间区间来估算的, 包括周转库存积压的资金费用, 由于零部件过期的价格损失, 以及由戴尔评估的库存费. 惩罚费用 p 由两项因素估算: 取消订单或订单缺货所引起的利润损失, 以及需要时零部件供应不上所增加的运输费用.

---

① G. Hadley and T. Whitin, *Analysis of Inventory Systems*, Prentice Hall, Englewood Cliffs, NJ, 1963, p.153.

提前时间期间的需求量可以根据若干因素来估算：(1) 戴尔公司从周转库存提出的每日消费量 (提货量); (2) 供应商对周转库存的每日供货量; (3) 从供应商到周转库存的日运输量; (4) 使用 XDX 零部件的 PC 日销售量; (5) 戴尔给供应商关于下个月预期使用 XDX 零部件的预报; (6) 供应商的提前时间.

数据分析

数据分析的第一个任务是, 评价这 3 家 XDX 供应商的库存策略按照上面给出的 (R,Q) 策略是否已经最优. 我们通过计算在数据采集期间 (1998 年 12 月 1 日到 1999 年 5 月 27 日) 内的所有订单的 Z 值来检验这项判断是否成立. 具体来讲, 当给定实际的再订货水平 R、估计的 (日) 需求量平均值和方差以及提前时间 (天数) 后, 从上面给出的公式可以算出

$$Z = \frac{R - DT}{\sqrt{D^2\sigma_t^2 + T\sigma_d^2}}$$

计算结果表明, 在考察期间上的 Z 值变化很大, 从 0.8 到 5.2. 假设估计的储存费用 h 和惩罚费用 p 对所有 3 家供应商相同, 最优条件下的 Z 值应该保持不变, 约为 $Z = \Phi^{-1}(\frac{p}{p+h})$. 这些供应商的订货策略差别非常大 (从 0.8 到 5.2), 这一事实就说明了他们的库存策略不是最优的, 这就使得研究小组决定要找出背后的原因.

供应商按照戴尔提供的月需求预测来制定他的补充库存策略. 这些预测可以有 3 种水平：(1) **合计数**, 表示下个月内要生产的 PC 总数; (2)**配件数**, 表示要使用 XDX 部件的 PC 数; (3)**提货数**, 代表戴尔要从每家供应商周转库存中提出的 XDX 零部件数. 对合计数、配件数和提货数预测误差结果造成需求变化相应的增加, 用需求分布的方差来表示. 对这 3 种预测值的研究显示, 戴尔的预测误差造成周转总库存增加了将近 50%, 主要是由于这些供应商想要通过保存更大的安全库存来应付较大的需求方差. 这个信息非常重要, 因为除了按照现有的“矛盾的”预测来实现最优的 (R,Q) 系统外, 戴尔在采用更加准确的预测技术来降低周转库存方面可以发挥重要的作用. 尤其是, 提货预测误差主要是由于戴尔一次从一家供应商那里提出所需要的 XDX 零部件, 而不是从所有供应商那里按照一定比例提货造成的. 从周转库存“成块地”提出 XDX, 造成了较高的补货率, 因此导致安全库存量过大.

结论

研究结果显示, 消除成块提货并提供更加可靠的预测, 有可能降低 38% 的安全库存量. 到目前为止, 戴尔一条生产线已经减低了 20% 的库存量, 预计每年节约 270 万美元. 这项研究中最重要的发现就是认识到, 可以通过消除由于预测误差引起的方差来源, 从而降低库存水平. 这项研究结果也符合最优再订货水平直接依赖于需求的标准差的事实.

思考题

1. 通过将平均库存水平与从所有供应商等量提货的库存水平相比较, 评价周转库存"成块"提货的反效果.
2. 假设一条组装线每天的平均需求量为 $N(300, 10)$, 提前时间期间的需求服从 $N(50, 4)$. 已知每天每件未满足的需求的惩罚费用是 \$100, 每天每件剩余库存的储存费用为 \$5, 求出再订货水平.
3. 用第 2 题给出的数据, 研究再订货水平对日平均需求量估计误差 $\pm r\%$ 的灵敏度 (r 的变化范围是 0~50). 标准差保持不变.

案例 14 某制造厂内部运输系统的分析[①]

工具: 排队论, 模拟

应用领域: 原材料供应

问题描述

有 3 辆卡车用于在一家制造厂内部运送原材料. 这些卡车都在中心停车场等待发车请求. 一辆收到发车请求的卡车要开到顾客地点, 装车后开到目的地, 然后再返回到中心停车场. 这项服务的主要用户包括生产部门 (P)、制造车间 (W) 和维修部门 (M). 其他部门 (O) 偶尔也会请求使用这些卡车. 用户对空车的等待时间过长提出了抱怨, 特别是生产部门, 他们要求车队增加第 4 辆卡车. 本研究是针对第 4 辆卡车所增加费用的合理性进行分析.

所需要的输入数据

我们采集了一段 17 天连续的两班倒工作日期间的该内部运输系统的有关运行信息. 表 24.21 和表 24.22 给出了所采集信息的汇总情况.

表 24.21 内部运输系统的运行数据汇总

	卡车用户				
	P	W	M	O	总体平均
平均每小时请求数	3.02	0.84	0.26	0.48	4.6
平均每次请求卡车使用时间 (分)	18	25	32	20	20.3
每次请求卡车时间的标准差 (分)	8	11	15	14	10.6
一次卡车请求的平均等待时间 (分)	9.2	9.4	9.2	8.4	9.0

在表 24.21 中, 我们有了平均请求率 (到达率)、卡车在用的平均时间 (服务时间), 以及一次请求的平均等待时间. 表 24.22 给出了整个考察期间作为请求数的函数的在用卡车数.

① **资料来源**: G. P. Cosmetatos, "The Value of Queuing Theory—A Case Study," *Interfaces*, Vol.9, No.3, pp.47-51, 1979.

表 24.22　作为请求数的函数的在用卡车数

	提出请求时在用的卡车数				合　计
	0	1	2	3	
请求数	862	28	167	115	1 172
总百分比	73.6	2.4	14.2	9.8	

问题的分析

对得出表 24.21 中信息的原始数据进行分析, 可以看出下列现象:

(1) 对使用卡车的请求是随机的, 可以用泊松分布来表示.

(2) 服务时间 (卡车从开向顾客到返回停车场之间的使用时间) 是单峰偏移的, 似乎不服从指数分布. 针对这种情况, 我们可以用三角分布来近似地表示这个分布.

(3) 虽然运行中卡车的分配没有优先级, 但卡车司机还是愿意承载较近的顾客.

表 24.22 中的数据反映出两个特征:

(1) 在 73.6% 的请求时, 所有 3 辆卡车都是空闲的.

(2) 只有 9.8% 的请求时, 所有 3 辆卡车都在使用.

由于到达是随机的, 可以用泊松分布来描述, 而且服务时间非指数分布, 因此最符合这个问题的排队模型是 $M/G/c/\infty/\infty$. 但是, 对 $M/G/c$ 模型的计算不方便. 因此, 我们决定采用一个等价的 $M/M/c$ 模型作为队列中等待时间的一个上界估计. 这样做的合理性在于, 指数服务时间在所有的指数分布中是“最随机的”, 因此会导致本问题的一种最坏的情况. (按照同样的逻辑, 这个 $M/D/c$ 模型给出平均排队时间的一个下界, 因为服务时间是常数, 因此表达了这种“最不随机”的情况.)

下面总结该 $M/M/c$ 模型对 $c=3, \lambda=\frac{4.6}{60}=0.076\ 7$ 次请求/分钟, 以及 $\mu=\frac{1}{20.3}=0.049\ 3$ 次服务/分钟的结果:

系统空闲的概率 $p_0=0.197$

系统中至少有 3 个请求的概率 $p_{n\geqslant 3}=0.133$

平均队长 $L_q=0.277$ 个请求

队列平均等待时间 $W_q=3.6$ 分钟

观察这些结果, 我们注意到一个令人困惑的现象, 即 (从 $M/M/c$ 模型估算出来的) 队列平均等待时间的上界比实际观察到的要少得多 ($W_q=3.6$ 分钟, 与表 24.21 中给出的观察值 9.0 分钟相比). 从这一现象可以得出以下两个结论之一: 要么 λ 和 μ 的估计严重不准确, 要么平均等待时间的估计不可靠. 对数据的仔细研究表明, 这些数据的确是可靠的. 为了进一步证实 $M/M/c$ 模型的结果, 我们进行了模拟, 其中服务时间分布用参数为 $(15, 20.3, 30)$ 的三角分布近似表示. 中间的值代表所观察到的平均服务时间, 低值和高值根据服务时间的标准差 (=10.6 分钟) 和观察的最小和最大服务时间估计出来. 可以利用 Excel 程序 excelMultiServer.xls, 以每分钟 0.076 7 个请求的泊松到达率和三角服务时间来进行模拟. 用 10 次重复, 每次模拟

450 个请求, 发现平均排队时间的变化范围是, 最少 1.1 分钟, 最多 3.62 分钟, 平均值为 2.07 分钟. 这一结果使我们相信, 从 $M/M/c$ 模型得到的 3.6 分钟的上界结果是对的. 此外, 从观察数据 (=9.0 分钟) 得到的长等待时间似乎与表 24.22 中的数据互相矛盾, 表中当一个服务请求到达时, 有 73.6% 的时间所有 3 辆卡车都空闲.

那么如何来解决观察结果与估计结果的不一致性呢? 当回到工厂进一步研究这个运输系统时, 分析人员幸运地发现, 是停车场的布局有问题, 使得用户看不到正在等待的卡车, 因而他们就以为那里没有车了. 其实这就相当于运行的卡车不足 3 辆, 从而导致了人为增加的等待时间. 一旦发现了这个问题, 解决办法就很显然了: 即让卡车司机与用户之间有一种双向的通讯系统. 这样提出的解决方案一下子改善了服务, 等待时间大大减少.

虽然提出的解决方案并不是直接由排队分析结果"推出来"的, 但它确符合排队分析的逻辑本质, 即从中发现数据的不一致性, 并因此能抓住问题的源头.

思考题

1. 在表 24.21 中, "总体平均"列的各值是如何算出来的?
2. 若一辆卡车 95% 的服务请求都来自生产部门, 只有 5% 来自其他部门, 对这种情况你会如何分析.
3. 通过研究模拟结果对三角分布最小值和最大值的变化的灵敏度, 来评价分析结果的稳健性. 这项灵敏度分析是会肯定还是会否定本案例给出的结论?

案例 15 Qantas 航空公司电话售票人力资源计划问题①

工具: ILP, 排队论

应用领域: 航空公司

问题描述

为了降低经营成本, Qantas 航空公司想要在本部电话订票办公室合理地安排工作人员, 以对顾客提供方便的服务. 订票处每天的办公时间是从早 7:00 到晚 22:00, 有 6 个班组从 7:00 到 9:30 之间的每整点和半点开始上班, 还有一个班从 15:00 开始上班. 从 7:00 到 8:01 之间上班的班组工作 8 个小时, 从 8:30 到 9:31 之间开始上班的班组工作 $8\frac{1}{2}$ 小时. 第 7 个班组从 15:00 开始, 只工作 7 个小时. 所有接线员有半个小时的休息时间, 但是从 8 : 30 到 9 : 31 之间开始上班的人休息时间有一个小时. 接线员至少工作 3 个小时以后才能休息, 两个小时内每半个小时安排一次休息时间. 这些上班时间和休息时间长度的安排, 是为了能够对 7 : 00 到 22 : 00 一天时间内的工作小时进行"合理"的分布.

① **资料来源**: Gaballa, A. and W. Pearce, "Telephone Sales Manpower Planning at Qantas," *Interfaces*, Vol.9, No.3, pp.1-9, May 1979.

按照传统方法, 对员工需求的估计是根据过去的业务增长情况来预测未来电话呼叫量. 需要增加的员工数就可以根据一个接线员能够处理的平均呼叫量去除预测的电话呼叫增长量计算出来. 由于是根据平均数进行的计算, 增加雇用的员工数并没有考虑到每天需求的变化情况. 特别是, 高峰业务时间对服务的等待时间过长导致了顾客的抱怨, 并失去了一些业务.

要解决这个问题, 就需要制定一个计划, 使得在雇用的接线员数与顾客需求之间达到某种平衡. 这个计划必须要考虑到营业时间内需求的波动变化.

数据采集

本研究所需的数据包括到达呼叫的分布和接线员接待顾客所花时间 (服务时间) 的分布. 这些数据收集来以后, 再按照 3 个月期间每半个小时长度一个记录. 数据分析得到以下主要结果:

(1) 一天内不同时间的到达量显示出很大变化, 但不同日期之间是一致的.

(2) 服务时间的分布基本上随时间平稳.

(3) 半个小时期间的到达分布近似于泊松分布, 虽然平均值不同 (见表 24.23).

(4) 服务时间分布近似于平均值为 3.5 分钟的指数分布.

表 24.23 7:00 到 22:00 期间每 30 分钟时间段计算得出的每分钟平均呼叫数

时段	均值	时段	均值
7:00～7:30	5	14:30～15:00	17
7:30～8:00	8	15:00～15:30	15
8:00～8:30	10	15:30～16:00	12
8:30～9:00	10	16:00～16:30	10
9:00～9:30	12	16:30～17:00	10
9:30～10:00	12	17:00～17:30	10
10:00～10:30	15	17:30～18:00	9
10:30～11:00	17	18:00～18:30	9
11:00～11:30	20	18:30～19:00	9
11:30～12:00	23	19:00～19:30	7
12:00～12:30	23	19:30～20:00	7
12:30～13:00	23	20:00～20:30	5
13:00～13:30	21	20:30～21:00	5
13:30～14:00	17	21:00～21:30	3
14:00～14:30	17	21:30～22:00	3
平均服务时间 = 3.5 分钟			

数学模型

可以用一个整数线性规划来描述这个问题, 目标是在满足全天半小时需求波动以及满足班组工作时间和休息时间的条件下, 使得雇用接线员的总数达到最少. 令

$x_j=$ 班组 j 开始上班的接线员数, $j=1,2,\cdots,7$

$y_{ij}=$ 在第 i 个半小时期间休息的来自班组 j 的接线员数, $i=1,2,3,4$

$c_k=$ 第 k 个半小时内需要的最少的接线员数, $k=1,2,\cdots,30$

因此给出的模型为:

$$
\begin{aligned}
\min\quad & z=\sum_{j=1}^{7}x_j \\
\text{s.t.}\quad & \sum_{r=1}^{k}x_r \geqslant c_k,\quad k=1,2,\cdots,6 \\
& \sum_{r=1}^{6}x_r - y_{11}\geqslant c_7 \\
& \sum_{r=1}^{6}x_r - y_{21} - y_{12}\geqslant c_8 \\
& \sum_{r=1}^{6}x_r - y_{31} - y_{22} - y_{13}\geqslant c_9 \\
& \sum_{r=1}^{6}x_r - y_{41} - y_{32} - y_{23} - y_{14}\geqslant c_{10} \\
& \sum_{r=1}^{6}x_r - y_{42} - y_{33} - y_{14} - y_{24} - y_{15}\geqslant c_{11} \\
& \sum_{r=1}^{6}x_r - y_{43} - y_{24} - y_{34} - y_{15} - y_{25} - y_{16}\geqslant c_{12} \\
& \sum_{r=1}^{6}x_r - y_{34} - y_{44} - y_{25} - y_{35} - y_{16} - y_{26}\geqslant c_{13} \\
& \sum_{r=1}^{6}x_r - y_{44} - y_{35} - y_{45} - y_{26} - y_{36}\geqslant c_{14} \\
& \sum_{r=1}^{6}x_r - y_{45} - y_{36} - y_{46}\geqslant c_{15} \\
& \sum_{r=1}^{6}x_r - y_{46}\geqslant c_{16} \\
& x_2+x_3+x_4+x_5+x_6+x_7\geqslant c_{17} \\
& x_3+x_4+x_5+x_6+x_7\geqslant c_{18} \\
& x_4+x_5+x_6+x_7\geqslant c_{19} \\
& x_4+x_5+x_6+x_7\geqslant c_{20} \\
& x_5+x_6+x_7-y_{17}\geqslant c_{21} \\
& x_6+x_7-y_{27}\geqslant c_{22} \\
& x_7-y_{37}\geqslant c_{23}
\end{aligned}
$$

$$x_7 - y_{47} \geqslant c_{23}$$

$$x_7 \geqslant c_k,\ k = 25, 26, \cdots, 30$$

$$\sum_{i=1}^{4} y_{ij} - x_j = 0,\ j = 1, 2, \cdots, 7$$

所有 x_j, y_{ij} 为非负整数

该模型约束的逻辑含义是: 对于开始时间为 7 : 00, 7 : 30, 8 : 00 的第 1, 2, 3 班, 接线员工作 8 小时, 等于 16 个半小时段. 这就是为什么 x_1, x_2, x_3 每个变量连续出现在 16 个约束里的原因. 按照同样的逻辑, 变量 x_4, x_5, x_6 对应于从 8 : 30, 9 : 00, 9 : 30 开始上班的班组, 每班工作 $8\frac{1}{2}$ 小时. 因此, 每个这些变量连续出现在 17 个约束里. 最后, x_7 代表从 15 : 00 开始上班的班组 (时段 17), 工作 7 小时, 因此从时段 17 开始, 连续出现在 14 个约束中. 对于表示休息时间的变量 y_{ij}, 有两种情况来说明约束逻辑: 对于第 1 班 (7:00 开始上班), 有 4 次半小时休息时段, 从 10:00(时段 7)、10:30(时段 8)、11:00(时段 9)、11:30(时段 10) 开始. 这就是为什么从约束 7 到 10 的左端减去 $y_{11}, y_{21}, y_{31}, y_{41}$ 的原因. 类似的原因用在第 2 班和第 3 班的 y_{ij}. 还有一种不同的情况是第 4 班, 从 8:30 开始, 持续 $8\frac{1}{2}$ 小时, 它的休息时段从 11:30(时段 10)、12:00(时段 11)、12:30(时段 12) 和 13:00(时段 13) 开始, 分别用变量 $y_{14}, y_{24}, y_{34}, y_{44}$ 表示. 因为这个班可以有一个小时休息时间, 所以这些变量的每一个都连续出现在两个约束中. 例如, y_{14} 出现在约束 10 和约束 11 里. 最后一个约束保证每个接线员都会有一个中间休息时段.

建立上述模型所需要的数据就是约束的右端项 $c_k, k = 1, 2, \cdots, 30$. 下面的部分说明如何用排队模型来求出这些值.

估计时段 k 所需要的最少接线员数 c_k

收集到的数据表明, 每半个小时时间段内到来的呼叫流量可以适当地用多服务台泊松排队模型 $(M/M/c):(GD/\infty/\infty)$ 表示. 每半小时内的呼叫到达率可以从表 24.23 给出的数据求出. 这些数据也表明, 服务时间的分布近似服从平稳均值 3.5 分钟的指数分布. 设 λ_k 为第 k 个半小时时段每分钟的到达率, 且 μ 是每分钟的服务率, 并定义 $[v]$ 为 $\leqslant v$ 的最大整数, 则时段 k 期间所需要的接线员数的一个下界为

$$c_k = \left[\frac{\lambda_k}{\mu}\right] + 1, k = 1, 2, \cdots, 30$$

例如, 对每分钟 $\lambda_1 = 5$ 个呼叫, 并且每分钟 $\mu = \frac{1}{3.5} = 0.286$ 次服务, 从 7:00 到 7:30 最少的接线员数为 $\left[\frac{5}{0.286}\right] + 1 = 17 + 1 = 18$. 这些计算背后的合理性在于, c_k 是维持这个排队模型中稳态条件所需要的最小值.

c_k 的下界估计并不提供关于对呼叫顾客服务质量的信息. 尤其是为了保证不丢失呼叫 (以及信誉损失), 接线员回答一次呼叫前的顾客等待时间应该充分短. 管

理部门设定了这方面的目标, 至少 90% 的呼叫应该在 20 秒之内得到回应. 这一目标写成下面的概率语句就是:

$$\mathrm{P}\{\text{等待时间}, t > T = 20\text{秒}\} < 0.1$$

通过设定不同的服务台数 c, 可以实现所需要的概率. 根据排队论的结果, 若 $\rho = \frac{\lambda}{\mu}$, 我们有①

$$\mathrm{P}\{t \leqslant T\} = \begin{cases} 1 - \dfrac{cp^c p_0}{c!(c-p)}, & T = 0 \\ \dfrac{p^c(1 - \mathrm{e}^{-\mu(c-p)T})p_0}{(c-1)!(c-p)} + \mathrm{P}\{t = 0\}, & T > 0 \end{cases}$$

12.6.3 节给出了求概率 p_0 的公式. 在时段 k 内, 所需要的最少接线员数的估计就是满足

$$\mathrm{P}\{t > T\} = 1 - \mathrm{P}\{t \leqslant T\} < 0.1$$

的最小的 c.

这条概率语句的性质决定了得不出闭形式的 c 值解. 相反, 通过替代 c 的连续增加值直到首次满足所需要的条件, 以得出需要的解. c 的下界 $(= \lfloor\rho\rfloor + 1)$ 为这种替代提供一个初始点.

我们专门设计了一个电子表格程序 (文件 excelCase15.xls), 非常适合用来进行这种计算, 计算结果见表 24.24.

AMPL 模型

文件 amplCase15.txt 是一个求解本问题的 AMPL 模型, 图 24.23 是该模型的解, 按照便于理解的方式重新做了安排. 该表显示, 总共需要 134 名接线员, 并详细给出了每个班组的接线员数以及指定的休息时间安排.

思考题

1. 关于当前要求 $8\frac{1}{2}$ 小时工作时间以及 1 个小时休息的班组 4,5,6(分别从 8:30, 9:00, 9:30 开始上班), 有一个有意思的问题. 大多数员工愿意工作时间短一些 (8 小时), 宁可只要半个小时的休息, 而不愿意有整个一个小时休息但是工作时间较长 ($8\frac{1}{2}$ 小时). 请研究一下, 假如对所有员工统一要求 8 小时工作、半小时休息, 这样会有什么结果.
2. 研究解对于应答一次呼叫中有 20 秒钟迟后的概率的变化的灵敏度.

---

① 见 D. Gross and C. Harris, *Fundamentals of Queuing Theory*, Wiley, New York, 1974, 公式 3.20, p.101. 也见习题15.6E 中的第 16 题.

表 24.24 最少接线员数的估计

时段 k	λ	c_k	P{$t > 20$ 秒}	时段 k	λ	c_k	P{$t > 20$ 秒}
1	5	23	0.090	16	17	74	0.082
2	8	35	0.074	17	15	63	0.090
3	10	42	0.093	18	12	50	0.077
4	10	42	0.093	19	10	42	0.093
5	12	50	0.077	20	10	42	0.093
6	12	50	0.077	21	10	42	0.093
7	15	63	0.090	22	9	39	0.068
8	17	74	0.082	23	9	39	0.068
9	18	79	0.090	24	9	39	0.068
10	19	85	0.095	25	7	31	0.080
11	20	90	0.095	26	7	31	0.080
12	20	90	0.095	27	5	23	0.090
13	20	90	0.095	28	5	23	0.090
14	19	85	0.082	29	3	15	0.092
15	18	79	0.082	30	3	15	0.092

```
HIRING SCHEDULE:
-------------------------------------
 Shift starting time      Shift size
-------------------------------------
        7:00                  26
        7:30                  21
        8:00                  25
        9:00                   4
        9:30                  19
       14:00                  39
-------------------------------------
                  Total =   134

BREAK SCHEDULE:
--------------------------------------------------------------------------
 Shift start time  Number of operators  Break start time  break time (min)
--------------------------------------------------------------------------
      7:00                26                 10:00               30
      7:30                21                 10:30               30
      8:00                15                 11:00               30
      8:00                10                 11:30               30
      9:00                 4                 12:00               60
      9:30                 4                 13:00               60
      9:30                15                 14:00               60
     14:00                20                 17:00               30
     14:00                19                 17:30               30
--------------------------------------------------------------------------
```

图 24.23 第 4 班组到第 6 班组工作 $8\frac{1}{2}$ 小时所需的员工数

附录B 统计表[①]

表 B.1 正态分布函数

$$F(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} \mathrm{e}^{-(\frac{1}{2})t^2}\mathrm{d}t$$

z	0.000	0.010	0.020	0.030	0.040	0.050	0.060	0.070	0.080	0.090
0.0	0.500 0	0.504 0	0.508 0	0.512 0	0.516 0	0.519 9	0.523 9	0.527 9	0.531 9	0.535 9
0.1	0.539 8	0.543 8	0.547 8	0.551 7	0.555 7	0.559 6	0.563 6	0.567 5	0.571 4	0.575 3
0.2	0.579 3	0.583 2	0.587 1	0.591 0	0.594 8	0.598 7	0.602 6	0.606 4	0.610 3	0.614 1
0.3	0.617 9	0.621 7	0.625 5	0.629 3	0.633 1	0.636 8	0.640 6	0.644 3	0.648 0	0.651 7
0.4	0.655 4	0.659 1	0.662 8	0.666 4	0.670 0	0.673 6	0.677 2	0.680 8	0.684 4	0.687 9
0.5	0.691 5	0.695 0	0.698 5	0.701 9	0.705 4	0.708 8	0.712 3	0.715 7	0.719 0	0.722 4
0.6	0.725 7	0.729 1	0.732 4	0.735 7	0.738 9	0.742 2	0.745 4	0.748 6	0.751 7	0.754 9
0.7	0.758 0	0.761 1	0.764 2	0.767 3	0.770 4	0.773 4	0.776 4	0.779 4	0.782 3	0.785 2
0.8	0.788 1	0.791 0	0.793 9	0.796 7	0.799 5	0.802 3	0.805 1	0.807 8	0.810 6	0.813 3
0.9	0.815 9	0.818 6	0.821 2	0.823 8	0.826 4	0.828 9	0.831 5	0.834 0	0.836 5	0.838 9
1.0	0.841 3	0.843 8	0.846 1	0.848 5	0.850 8	0.853 1	0.855 4	0.857 7	0.859 9	0.862 1
1.1	0.864 3	0.866 5	0.868 6	0.870 8	0.872 9	0.874 9	0.877 0	0.879 0	0.881 0	0.883 0
1.2	0.884 9	0.886 9	0.888 8	0.890 7	0.892 5	0.894 4	0.896 2	0.898 0	0.899 7	0.901 5
1.3	0.903 2	0.904 9	0.906 6	0.908 2	0.909 9	0.911 5	0.913 1	0.914 7	0.916 2	0.917 7
1.4	0.919 2	0.920 7	0.922 2	0.923 6	0.925 1	0.926 5	0.927 9	0.929 2	0.930 6	0.931 9
1.5	0.933 2	0.934 5	0.935 7	0.937 0	0.938 2	0.939 4	0.940 6	0.941 8	0.942 9	0.944 1
1.6	0.945 2	0.946 3	0.947 4	0.948 4	0.949 5	0.950 5	0.951 5	0.952 5	0.953 5	0.954 5
1.7	0.955 4	0.956 4	0.957 3	0.958 2	0.959 1	0.959 9	0.960 8	0.961 6	0.962 5	0.963 3
1.8	0.964 1	0.964 9	0.965 6	0.966 4	0.967 1	0.967 8	0.968 6	0.969 3	0.969 9	0.970 6
1.9	0.971 3	0.971 9	0.972 6	0.973 2	0.973 8	0.974 4	0.975 0	0.975 6	0.976 1	0.976 7
2.0	0.977 2	0.977 8	0.978 3	0.978 8	0.979 3	0.979 8	0.980 3	0.980 8	0.981 2	0.981 7
2.1	0.982 1	0.982 6	0.983 0	0.983 4	0.983 8	0.984 2	0.984 6	0.985 0	0.985 4	0.985 7
2.2	0.986 1	0.986 4	0.986 8	0.987 1	0.987 5	0.987 8	0.988 1	0.988 4	0.988 7	0.989 0
2.3	0.989 3	0.989 6	0.989 8	0.990 1	0.990 4	0.990 6	0.990 9	0.991 1	0.991 3	0.991 6
2.4	0.991 8	0.992 0	0.992 2	0.992 5	0.992 7	0.992 9	0.993 1	0.993 2	0.993 4	0.993 6
2.5	0.993 8	0.994 0	0.994 1	0.994 3	0.994 5	0.994 6	0.994 8	0.994 9	0.995 1	0.995 2
2.6	0.995 3	0.995 5	0.995 6	0.995 7	0.995 9	0.996 0	0.996 1	0.996 2	0.996 3	0.996 4
2.7	0.996 5	0.996 6	0.996 7	0.996 8	0.996 9	0.997 0	0.997 1	0.997 2	0.997 3	0.997 4
2.8	0.997 4	0.997 5	0.997 6	0.997 7	0.997 7	0.997 8	0.997 9	0.997 9	0.998 0	0.998 1
2.9	0.998 1	0.998 2	0.998 2	0.998 3	0.998 4	0.998 4	0.998 5	0.998 5	0.998 6	0.998 6
3.0	0.998 7	0.998 7	0.998 7	0.998 8	0.998 8	0.998 9	0.998 9	0.998 9	0.999 0	0.999 0
3.1	0.999 0	0.999 1	0.999 1	0.999 1	0.999 2	0.999 2	0.999 2	0.999 2	0.999 3	0.999 3
3.2	0.999 3	0.999 3	0.999 4	0.999 4	0.999 4	0.999 4	0.999 4	0.999 5	0.999 5	0.999 5
3.3	0.999 5	0.999 5	0.999 5	0.999 6	0.999 6	0.999 6	0.999 6	0.999 6	0.999 6	0.999 7
3.4	0.999 7	0.999 7	0.999 7	0.999 7	0.999 7	0.999 7	0.999 7	0.999 7	0.999 7	0.999 8

① Excel 电子表格 excelStartTable.xls 可以取代 (硬拷贝的) 12 种常见分布的统计表, 也包括出现在本附录中的表格.

(续)

z	0.000	0.010	0.020	0.030	0.040	0.050	0.060	0.070	0.080	0.090
3.5	0.999 8									
4.0	0.999 97									
5.0	0.999 999 7									
6.0	0.999 999 999									

资料来源: Miller, I., and J. Freund, *Probability and Statistics for Engineers*, Prentice Hall, Upper Saddle River, NJ, 1985.

表 B.2 $t_{\alpha,\nu}$ 值 *

ν	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$	ν
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.131	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
Inf	1.282	1.645	1.960	2.326	2.576	Inf

* 摘自 Macmillan Publishing Co., Inc., *Statistical Methods for Research Workers*, 14th ed, R. A. Fisher. Copyright © 1970 University of Adelaide, 并经许可.

表 B.3 $\chi^2_{\alpha,\nu}$ 值 *

ν	$\alpha=0.995$	$\alpha=0.99$	$\alpha=0.975$	$\alpha=0.95$	$\alpha=0.05$	$\alpha=0.025$	$\alpha=0.01$	$\alpha=0.005$	ν
1	0.000 039 3	0.000 157	0.000 982	0.003 93	3.841	5.024	6.635	7.879	1
2	0.010 0	0.020 1	0.050 6	0.103	5.991	7.378	9.210	10.597	2
3	0.071 7	0.115	0.216	0.352	7.815	9.348	11.345	12.838	3
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860	4
5	0.412	0.554	0.831	1.145	11.070	12.832	15.056	16.750	5
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548	6
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278	7
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955	8
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589	9
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188	10
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757	11
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300	12
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819	13
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319	14
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801	15
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267	16
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718	17
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156	18
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582	19
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997	20
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401	21
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796	22
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181	23
24	9.886	10.856	12.401	13.484	36.415	39.364	42.980	45.558	24
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928	25
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290	26
27	11.808	12.879	14.573	16.151	40.113	43.194	46.963	49.645	27
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993	28
29	13.121	14.256	16.047	17.708	42.557	45.772	49.588	52.336	29
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672	30

* 本表基于*Biometrika Tables for Statisticians*, Vol. 1 的表 8, 并经 Biometrika 编委会许可.

附录 D　向量和矩阵复习

D.1　向　　量

D.1.1　向量的定义

令 $p_1, p_2, \cdots, p_n$ 是任意 n 个实数, 且 $\boldsymbol{P}$ 为这些实数的一个有序集合, 即

$$\boldsymbol{P} = (p_1, p_2, \cdots, p_n)$$

则称 $\boldsymbol{P}$ 是一个 n 维向量 (或简称为向量), $\boldsymbol{P}$ 的第 i 个元素就是 p_i. 例如, $\boldsymbol{P} = (1, 2)$ 为一个二维向量.

D.1.2　向量的相加 (相减)

考虑 n 维向量

$$\begin{aligned} \boldsymbol{P} &= (p_1, p_2, \cdots, p_n) \\ \boldsymbol{Q} &= (q_1, q_2, \cdots, q_n) \\ \boldsymbol{R} &= (r_1, r_2, \cdots, r_n) \end{aligned}$$

向量 $\boldsymbol{R} = \boldsymbol{P} \pm \boldsymbol{Q}$ 的第 i 个元素为 $r_i = p_i \pm q_i$. 一般地, 若已知向量 $\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{S}$, 则

$$\begin{array}{ll} \boldsymbol{P} + \boldsymbol{Q} = \boldsymbol{Q} + \boldsymbol{P} & (\text{交换律}) \\ (\boldsymbol{P} \pm \boldsymbol{Q}) \pm \boldsymbol{S} = \boldsymbol{P} \pm (\boldsymbol{Q} \pm \boldsymbol{S}) & (\text{结合律}) \\ \boldsymbol{P} + (-\boldsymbol{P}) = \boldsymbol{0} & (\text{零向量}) \end{array}$$

D.1.3　标量与向量的乘积

已知向量 $\boldsymbol{P}$ 和标量 (常量)θ, 新的向量

$$\boldsymbol{Q} = \theta \boldsymbol{P} = (\theta p_1, \theta p_2, \cdots, \theta p_n)$$

是 $\boldsymbol{P}$ 与 θ 的标量积. 一般地, 给定向量 $\boldsymbol{P}$, $\boldsymbol{S}$ 和标量 θ, γ,

$$\begin{array}{ll} \theta(\boldsymbol{P} + \boldsymbol{S}) = \theta \boldsymbol{P} + \theta \boldsymbol{S} & (\text{分配律}) \\ \theta(\gamma \boldsymbol{P}) = (\theta\gamma) \boldsymbol{P} & (\text{结合律}) \end{array}$$

D.1.4　线性无关向量

向量 $\boldsymbol{P}_1, \boldsymbol{P}_2, \cdots, \boldsymbol{P}_n$ 是线性无关的, 当且仅当

$$\sum_{j=1}^{n} \theta_j \boldsymbol{P}_j = 0 \Rightarrow \theta_j = 0, \quad j = 1, 2, \cdots, n$$

若

$$\sum_{j=1}^{n}\theta_j \boldsymbol{P}_j = 0, \quad \text{对于某些}\ \theta_j \neq 0$$

则这些向量是*线性相关的*. 例如, 向量

$$\boldsymbol{P}_1 = (1,2),\ \boldsymbol{P}_2 = (2,4)$$

是线性相关的, 因为对于 $\theta_1 = 2$ 和 $\theta_2 = -1$, 有

$$\theta_1 \boldsymbol{P}_1 + \theta_2 \boldsymbol{P}_2 = 0$$

D.2 矩 阵

D.2.1 矩阵的定义

矩阵是各个元素的一个矩形数组. 矩阵 $\boldsymbol{A}$ 的元素 a_{ij} 位于该数组的第 i 行第 j 列. 一个有 m 行和 n 列的矩阵称为 $m \times n$ 大小 (阶) 的矩阵. 例如, 下面的矩阵是一个 (4×3) 阶的矩阵:

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{pmatrix} = \|a_{ij}\|_{4\times 3}$$

D.2.2 各种类型的矩阵

(1) 对于*正方*矩阵, 有 $m = n$.

(2) 单位矩阵是一个正方矩阵, 其中主对角元素等于 1, 而且所有非对角元素都等于 0. 例如一个 (3×3) 的单位矩阵为

$$\boldsymbol{I}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

(3) *行矩阵*是一个 1 行 n 列的矩阵.

(4) *列矩阵*是一个 m 行 1 列的矩阵.

(5) 矩阵 $\boldsymbol{A}^{\mathrm{T}}$ 称为矩阵 $\boldsymbol{A}$ 的转置(transpose), 若对于所有的 i 和 j, $\boldsymbol{A}$ 中的元素 a_{ij} 等于 $\boldsymbol{A}^{\mathrm{T}}$ 中的元素 a_{ji}. 例如,

$$\boldsymbol{A} = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} \Rightarrow \boldsymbol{A}^{\mathrm{T}} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

(6) 矩阵 $\boldsymbol{B} = \boldsymbol{0}$ 是一个零矩阵(zero matrix), 若 $\boldsymbol{B}$ 的所有元素都等于零.

(7) 两个矩阵 $\boldsymbol{A} = \|a_{ij}\|$, $\boldsymbol{B} = \|b_{ij}\|$ 相等, 当且仅当它们的阶相同, 并且对于所有的 i, j 都有 $a_{ij} = b_{ij}$.

D.2.3 矩阵的代数运算

矩阵中只定义了加法 (减法) 和乘法. 虽然没有定义除法, 但可以用求逆来代替它 (见附录 D.2.6).

矩阵的加法 (减法) 对于两个矩阵 $\boldsymbol{A}=\|a_{ij}\|$ 和 $\boldsymbol{B}=\|b_{ij}\|$, 若它们同样都是 $(m\times n)$ 阶的, 则可以相加, 和矩阵 $\boldsymbol{D}=\boldsymbol{A}+\boldsymbol{B}$ 由对应元素相加得到. 因此有

$$\|d_{ij}\|_{m\times n}=\|a_{ij}+b_{ij}\|_{m\times n}$$

若矩阵 $\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}$ 的阶相同, 则

$$\begin{aligned}&\boldsymbol{A}+\boldsymbol{B}=\boldsymbol{B}+\boldsymbol{A} && \text{(交换律)}\\ &\boldsymbol{A}\pm(\boldsymbol{B}\pm\boldsymbol{C})=(\boldsymbol{A}\pm\boldsymbol{B})\pm\boldsymbol{C} && \text{(结合律)}\\ &(\boldsymbol{A}\pm\boldsymbol{B})^{\mathrm{T}}=\boldsymbol{A}^{\mathrm{T}}\pm\boldsymbol{B}^{\mathrm{T}}\end{aligned}$$

矩阵的乘积 对于两个矩阵 $\boldsymbol{A}=\|a_{ij}\|$ 和 $\boldsymbol{B}=\|b_{ij}\|$, 当且仅当 $\boldsymbol{A}$ 的列数等于 $\boldsymbol{B}$ 的行数, 它们的乘积 $\boldsymbol{D}=\boldsymbol{AB}$ 有定义. 若 $\boldsymbol{A}$ 是 $(m\times r)$ 阶矩阵, $\boldsymbol{B}$ 是 $(r\times n)$ 阶矩阵, 则 $\boldsymbol{D}$ 是 $(m\times n)$ 矩阵, 其中 m 和 n 为任意正整数值. 在这种情况下, $\boldsymbol{D}$ 的元素计算为

$$d_{ij}=\sum_{k=1}^{r}a_{ik}b_{kj},\quad \text{对于所有的 } i \text{ 和 } j$$

例如, 若

$$\boldsymbol{A}=\begin{pmatrix}1&3\\2&4\end{pmatrix},\quad \boldsymbol{B}=\begin{pmatrix}5&7&9\\6&8&0\end{pmatrix}$$

我们有

$$\begin{aligned}\boldsymbol{D}&=\begin{pmatrix}1&3\\2&4\end{pmatrix}\begin{pmatrix}5&7&9\\6&8&0\end{pmatrix}=\begin{pmatrix}1\times5+3\times6&1\times7+3\times8&1\times9+3\times0\\2\times5+4\times6&2\times7+4\times8&2\times9+4\times0\end{pmatrix}\\&=\begin{pmatrix}23&31&9\\34&46&18\end{pmatrix}\end{aligned}$$

一般情况下, $\boldsymbol{AB}\neq\boldsymbol{BA}$, 即使 $\boldsymbol{BA}$ 有定义.

矩阵乘法服从下列一般性质:

$$\begin{aligned}&\boldsymbol{I}_m\boldsymbol{A}=\boldsymbol{A}\boldsymbol{I}_n=\boldsymbol{A},\quad \boldsymbol{I}_m,\ \boldsymbol{I}_n \text{ 为单位矩阵}\\&(\boldsymbol{AB})\boldsymbol{C}=\boldsymbol{A}(\boldsymbol{BC})\\&\boldsymbol{C}(\boldsymbol{A}\pm\boldsymbol{B})=\boldsymbol{CA}\pm\boldsymbol{CB}\\&(\boldsymbol{A}\pm\boldsymbol{B})\boldsymbol{C}=\boldsymbol{AC}\pm\boldsymbol{BC}\\&\alpha\boldsymbol{AB}=(\alpha\boldsymbol{A})\boldsymbol{B}=\boldsymbol{A}(\alpha\boldsymbol{B}),\ \alpha \text{ 是一个标量}\end{aligned}$$

分块矩阵的乘积 令 $\boldsymbol{A}$ 是一个 $(m\times r)$ 阶矩阵, $\boldsymbol{B}$ 是一个 $(r\times n)$ 阶矩阵, 设 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 按如下分块:

$$\boldsymbol{A}=\begin{pmatrix}\boldsymbol{A}_{11}&\boldsymbol{A}_{12}&\boldsymbol{A}_{13}\\\boldsymbol{A}_{21}&\boldsymbol{A}_{22}&\boldsymbol{A}_{23}\end{pmatrix},\quad \boldsymbol{B}=\begin{pmatrix}\boldsymbol{B}_{11}&\boldsymbol{B}_{12}\\\boldsymbol{B}_{21}&\boldsymbol{B}_{22}\\\boldsymbol{B}_{31}&\boldsymbol{B}_{32}\end{pmatrix}$$

分块中假定, 对于所有的 i, j, A_{ij} 的列数等于 B_{ij} 的行数, 则

$$A \times B = \begin{pmatrix} A_{11}B_{11} + A_{12}B_{12} + A_{13}B_{31} & A_{11}B_{12} + A_{12}B_{22} + A_{13}B_{32} \\ A_{21}B_{11} + A_{22}B_{21} + A_{23}B_{31} & A_{21}B_{12} + A_{22}B_{22} + A_{23}B_{32} \end{pmatrix}$$

例如,

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 0 & 5 \\ 2 & 5 & 6 \end{pmatrix} \begin{pmatrix} 4 \\ 1 \\ 8 \end{pmatrix} = \begin{pmatrix} (1)(4) + (2 \quad 3)\begin{pmatrix} 1 \\ 8 \end{pmatrix} \\ \hline \begin{pmatrix} 1 \\ 2 \end{pmatrix}(4) + \begin{pmatrix} 0 & 5 \\ 5 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 8 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} 4+2+24 \\ \hline \begin{pmatrix} 4 \\ 8 \end{pmatrix} + \begin{pmatrix} 40 \\ 53 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} 30 \\ 44 \\ 61 \end{pmatrix}$$

D.2.4 正方矩阵的行列式

考虑 n 阶正方矩阵

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

接下来, 定义乘积

$$P_{j_1 j_2 \cdots j_n} = a_{1j_1} a_{2j_2} \cdots a_{nj_n}$$

使得 A 的每列和每行在下标 $j_1, j_2, \cdots, j_n$ 中恰好被表示一次. 再定义

$$\epsilon_{j_1 j_2 \cdots j_n} = \begin{cases} 1, & j_1 j_2 \cdots j_n \text{是偶排列} \\ 0, & j_1 j_2 \cdots j_n \text{是奇排列} \end{cases}$$

令 ρ 表示所有 $n!$ 个排列的总和, 则 A 的行列式计算公式为

$$\sum_{\rho} \epsilon_{j_1 j_2 \cdots j_n} P_{j_1 j_2 \cdots j_n}$$

记为 $\det A$ 或 $|A|$.

设有矩阵

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

则

$$|A| = a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{31}a_{23}) + a_{13}(a_{21}a_{32} - a_{22}a_{31})$$

行列式的性质有以下几点.

(1) 若一行或一列中所有元素都为零, 则该行列式的值等于零.

(2) $|A| = |A^{\mathrm{T}}|$.

(3) 若 B 是由 A 经过交换任意两行或两列得到的, 则 $|B| = -|A|$.

(4) 若 A 的两行 (或两列) 相互成倍数关系, 则 $|A| = 0$.

(5) 若标量 α 乘以某一列 (行) 向量加到另一列向量上, 则 $|A|$ 的值不变.

(6) 若一个行列式的一列或一行的每个元素都乘以一个标量 α, 则新行列式的值等于原行列式的值乘以该标量.

(7) 若 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 是两个 n 阶方阵, 则

$$|\boldsymbol{AB}| = |\boldsymbol{A}||\boldsymbol{B}|$$

行列式子式的定义 行列式 $|\boldsymbol{A}|$ 中元素 a_{ij} 的子式 M_{ij} 是从矩阵 $\boldsymbol{A}$ 经过去掉第 i 行和第 j 列得到的. 例如, 对于

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

$$M_{11} = \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}, \quad M_{22} = \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix}, \cdots$$

伴随矩阵的定义 令 $A_{ij} = (-1)^{i+j} M_{ij}$ 定义为正方矩阵 $\boldsymbol{B}$ 的元素 a_{ij} 的**余子式**(cofactor), 则矩阵 $\boldsymbol{A}$ 的伴随矩阵即为 $||A_{ij}||$ 的转置, 定义为:

$$\text{adj } \boldsymbol{A} = \|A_{ij}\|^{\mathrm{T}} = \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \vdots & \vdots & & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{pmatrix}$$

例如, 若

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 2 \\ 3 & 3 & 4 \end{pmatrix}$$

则 $A_{11} = (-1)^2(3 \times 4 - 2 \times 3) = 6, A_{12} = (-1)^3(2 \times 4 - 3 \times 2) = -2, \cdots$, 或

$$\text{adj } \boldsymbol{A} = \begin{pmatrix} 6 & 1 & -5 \\ -2 & -5 & 4 \\ -3 & 3 & -1 \end{pmatrix}$$

D.2.5 非奇异矩阵

一个矩阵的秩是 r, 若该矩阵中有非零行列式的最大正方子矩阵的阶数为 r. 一个具有非零行列式的正方矩阵称为是**满秩** (full-rank) 或**非奇异** (nonsingular) 矩阵. 例如考虑矩阵

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 5 & 7 \end{pmatrix}$$

$\boldsymbol{A}$ 是**奇异** (singular) 矩阵, 因为

$$|\boldsymbol{A}| = 1 \times (21 - 20) - 2 \times (14 - 12) + 3 \times (10 - 9) = 0$$

但 $\boldsymbol{A}$ 的秩 $r = 2$, 因为

$$\begin{vmatrix} 1 & 2 \\ 2 & 3 \end{vmatrix} = -1 \neq 0$$

D.2.6 非奇异矩阵的逆矩阵

若 $\boldsymbol{B}$ 和 $\boldsymbol{C}$ 是两个 n 阶矩阵, 使得 $\boldsymbol{BC}=\boldsymbol{CB}=\boldsymbol{I}$, 则称 $\boldsymbol{B}$ 是 $\boldsymbol{C}$ 的逆, 而且 $\boldsymbol{C}$ 也是 $\boldsymbol{B}$ 的逆. 逆矩阵的常用记号是 $\boldsymbol{B}^{-1}$ 和 $\boldsymbol{C}^{-1}$.

定理 若 $\boldsymbol{BC}=\boldsymbol{I}$, 并且 $\boldsymbol{B}$ 是非奇异的, 则 $\boldsymbol{C}=\boldsymbol{B}^{-1}$, 这意味着逆矩阵是唯一的.

证明 由假设

$$\boldsymbol{BC}=\boldsymbol{I}$$

因此有

$$\boldsymbol{B}^{-1}\boldsymbol{BC}=\boldsymbol{B}^{-1}\boldsymbol{I}$$

得

$$\boldsymbol{IC}=\boldsymbol{B}^{-1}$$

即

$$\boldsymbol{C}=\boldsymbol{B}^{-1}$$

可以证明, 非奇异矩阵有以下两个重要的结果.

(1) 若 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 为非奇异 n 阶矩阵, 则 $(\boldsymbol{AB})^{-1}=\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.

(2) 若矩阵 $\boldsymbol{A}$ 非奇异, 则 $\boldsymbol{AB}=\boldsymbol{AC}$ 意味着 $\boldsymbol{B}=\boldsymbol{C}$.

矩阵的逆可用来求解 n 个线性无关方程. 考虑

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}=\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

其中 x_i 表示未知变量, a_{ij} 和 b_i 为常数. 这 n 个方程可以写成下面的矩阵形式:

$$\boldsymbol{AX}=\boldsymbol{b}$$

因为这些方程是独立的, 所以 $\boldsymbol{A}$ 必为非奇异. 因此有

$$\boldsymbol{A}^{-1}\boldsymbol{AX}=\boldsymbol{A}^{-1}\boldsymbol{b}$$

得

$$\boldsymbol{X}=\boldsymbol{A}^{-1}\boldsymbol{b}$$

D.2.7 矩阵求逆的计算方法[①]

伴随矩阵法 已知 $\boldsymbol{A}$ 是 n 阶非奇异矩阵,

$$\boldsymbol{A}^{-1}=\frac{1}{|\boldsymbol{A}|}\text{adj}\,\boldsymbol{A}=\frac{1}{|\boldsymbol{A}|}\begin{pmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \vdots & \vdots & & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{pmatrix}$$

① TORA 的求逆子模块基于 LU 分解方法. 见 Press 等人 (1986).

例如, 对于

$$A=\begin{pmatrix}1&2&3\\2&3&2\\3&3&4\end{pmatrix}$$

$$\text{adj}\ A=\begin{pmatrix}6&1&-5\\-2&-5&4\\-3&3&-1\end{pmatrix},\quad |A|=-7$$

因此有

$$A^{-1}=\frac{1}{-7}\begin{pmatrix}6&1&-5\\-2&-5&4\\-3&3&-1\end{pmatrix}=\begin{pmatrix}-\frac{6}{7}&-\frac{1}{7}&\frac{5}{7}\\\frac{2}{7}&\frac{5}{7}&-\frac{4}{7}\\\frac{3}{7}&-\frac{3}{7}&\frac{1}{7}\end{pmatrix}$$

行运算 (高斯–若尔当) 法 考虑分块矩阵 $(A|I)$, 其中 A 非奇异. 左乘 A^{-1}, 我们得到

$$(A^{-1}A|A^{-1}I)=(I|A^{-1})$$

因此, 应用特定的一系列行变换, 将 A 变成 I, I 变成 A^{-1}. 为了说明这一方法, 我们考虑方程组:

$$\begin{pmatrix}1&2&3\\2&3&2\\3&3&4\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}3\\4\\5\end{pmatrix}$$

X 的解和基矩阵的逆可直接从下式得到:

$$A^{-1}(A|I|b)=(I|A^{-1}|A^{-1}b)$$

我们给出变换运算的详细迭代步骤:

第 0 步

$$A=\left(\begin{array}{ccc|ccc|c}1&2&3&1&0&0&3\\2&3&2&0&1&0&4\\3&3&4&0&0&1&5\end{array}\right)$$

第 1 步

$$A=\left(\begin{array}{ccc|ccc|c}1&2&3&1&0&0&3\\0&-1&-4&-2&1&0&-2\\0&-3&-5&-3&0&1&-4\end{array}\right)$$

第 2 步

$$A=\left(\begin{array}{ccc|ccc|c}1&0&-5&-3&2&0&-1\\0&1&4&2&-1&0&2\\0&0&7&3&-3&1&2\end{array}\right)$$

第 3 步

$$A=\left(\begin{array}{ccc|ccc|c}1&0&0&-\frac{6}{7}&-\frac{1}{7}&\frac{5}{7}&\frac{3}{7}\\0&1&0&\frac{2}{7}&\frac{5}{7}&-\frac{4}{7}&\frac{6}{7}\\0&0&1&\frac{3}{7}&-\frac{3}{7}&\frac{1}{7}&\frac{2}{7}\end{array}\right)$$

我们得到 $x_1 = \frac{3}{7}$, $x_2 = \frac{6}{7}$, $x_3 = \frac{2}{7}$. $\boldsymbol{A}$ 的逆矩阵由矩阵的右端矩阵给出, 与伴随矩阵法的结果一致.

逆矩阵的乘积形式 假设我们有两个非奇异矩阵 $\boldsymbol{B}$ 和 $\boldsymbol{B}_{\text{next}}$, 这两个矩阵仅有一列不同. 进一步假设, $\boldsymbol{B}^{-1}$ 已知, 则逆矩阵 $\boldsymbol{B}_{\text{next}}^{-1}$ 可用下列公式计算:

$$\boldsymbol{B}_{\text{next}}^{-1} = \boldsymbol{E}\boldsymbol{B}^{-1}$$

矩阵 $\boldsymbol{E}$ 按如下方法计算: 若 $\boldsymbol{B}$ 中的列向量 $\boldsymbol{P}_j$ 用列向量 $\boldsymbol{P}_r$ 代替得到 $\boldsymbol{B}_{\text{next}}$, 则 $\boldsymbol{E}$ 的构造是把一个 m 阶单位矩阵的第 r 列用下列向量替换:

$$\xi = \frac{1}{(\boldsymbol{B}^{-1}\boldsymbol{P}_j)_r}\begin{pmatrix} -(\boldsymbol{B}^{-1}\boldsymbol{P}_j)_1 \\ -(\boldsymbol{B}^{-1}\boldsymbol{P}_j)_2 \\ \vdots \\ +1 \\ \vdots \\ -(\boldsymbol{B}^{-1}\boldsymbol{P}_j)_m \end{pmatrix} \leftarrow \text{第 } r \text{ 个位置}, \quad (\boldsymbol{B}^{-1}\boldsymbol{P}_j)_r \neq 0$$

若 $(\boldsymbol{B}^{-1}\boldsymbol{P}_j)_r = 0$, 则 $\boldsymbol{B}_{\text{next}}^{-1}$ 不存在.

公式 $\boldsymbol{B}_{\text{next}}^{-1}$ 的正确性证明如下. 定义 $\boldsymbol{F}$ 为一个 m 阶单位矩阵, 其第 r 列由 $(\boldsymbol{B}^{-1}\boldsymbol{P}_j)$ 代替, 即

$$\boldsymbol{F} = (\boldsymbol{e}_1, \cdots, \boldsymbol{e}_{r-1}, \boldsymbol{B}^{-1}\boldsymbol{P}_j, \boldsymbol{e}_{r+1}, \cdots, \boldsymbol{e}_m)$$

由于 $\boldsymbol{B}_{\text{next}}$ 与 $\boldsymbol{B}$ 的不同只有第 r 列由 $\boldsymbol{B}^{-1}\boldsymbol{P}_j$ 替换, 则

$$\boldsymbol{B}_{\text{next}} = \boldsymbol{B}\boldsymbol{F}$$

因此,

$$\boldsymbol{B}_{\text{next}}^{-1} = (\boldsymbol{B}\boldsymbol{F})^{-1} = \boldsymbol{F}^{-1}\boldsymbol{B}^{-1}$$

设 $\boldsymbol{E} = \boldsymbol{F}^{-1}$, 即得该公式.

这一乘积形式可用于按如下方法求出任何非奇异矩阵 $\boldsymbol{B}$ 的逆矩阵. 从 $\boldsymbol{B}_0 = \boldsymbol{I} = \boldsymbol{B}_0^{-1}$ 开始, 接下来用单位矩阵构造 $\boldsymbol{B}_1$, 其第一列用 $\boldsymbol{B}$ 的第一列代替, 则

$$\boldsymbol{B}_1^{-1} = \boldsymbol{E}_1\boldsymbol{B}_0^{-1} = \boldsymbol{E}_1\boldsymbol{I} = \boldsymbol{E}_1$$

一般地, 若用单位矩阵构造出 $\boldsymbol{B}_i$, 其第 i 列用矩阵 $\boldsymbol{B}$ 的第 i 列替换, 则有

$$\boldsymbol{B}_i^{-1} = \boldsymbol{E}_i\boldsymbol{B}_{i-1}^{-1} = \boldsymbol{E}_i\boldsymbol{E}_{i-1}\boldsymbol{B}_{i-2}^{-1} = \cdots = \boldsymbol{E}_i\boldsymbol{E}_{i-1}\cdots\boldsymbol{E}_1$$

这意味着对于初始矩阵 $\boldsymbol{B}$,

$$\boldsymbol{B}^{-1} = \boldsymbol{E}_n\boldsymbol{E}_{n-1}\cdots\boldsymbol{E}_1$$

下面的例子说明如何用乘积形式求矩阵的逆. 考虑

$$\boldsymbol{B} = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 4 & 0 & 1 \end{pmatrix}$$

第 0 步

$$B_0 = B_0^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

第 1 步

$$B_1 = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 4 & 0 & 1 \end{pmatrix}, \qquad B_0^{-1}P_1 = P_1 = \begin{pmatrix} 2 \\ 0 \\ 4 \end{pmatrix}^{\leftarrow r=1}$$

$$E_1 = \begin{pmatrix} +\frac{1}{2} & 0 & 0 \\ -\frac{0}{2} & 1 & 0 \\ -\frac{4}{2} & 0 & 1 \end{pmatrix}, \qquad B_1^{-1} = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix}$$

第 2 步

$$B_2 = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 4 & 0 & 1 \end{pmatrix} = B$$

$$B_1^{-1}P_2 = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 2 \\ -2 \end{pmatrix} \leftarrow r = 2$$

$$E_2 = \begin{pmatrix} 1 & -\frac{\frac{1}{2}}{2} & 0 \\ 0 & +\frac{1}{2} & 0 \\ 0 & -\frac{-2}{2} & 1 \end{pmatrix} = \begin{pmatrix} 1 & -\frac{1}{4} & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

$$B^{-1} = B_2^{-1} = E_2 B_1^{-1} = \begin{pmatrix} 1 & -\frac{1}{4} & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{4} & 0 \\ 0 & \frac{1}{2} & 0 \\ -2 & 1 & 1 \end{pmatrix}$$

分块矩阵法　假定有两个 n 阶非奇异矩阵 A 和 B, 按如下分块:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ (p \times p) & (p \times q) \\ A_{21} & A_{22} \\ (q \times p) & (q \times q) \end{pmatrix}, \quad A_{11}\text{非奇异}$$

$$B = \begin{pmatrix} B_{11} & B_{12} \\ (p \times p) & (p \times q) \\ B_{21} & B_{22} \\ (q \times p) & (q \times q) \end{pmatrix},$$

若 B 是 A 的逆矩阵, 则根据 $AB = I_n$, 我们有

$$A_{11}B_{11} + A_{12}B_{21} = I_p$$

$$A_{11}B_{12} + B_{12}A_{22} = 0$$

再根据 $\boldsymbol{BA}=\boldsymbol{I}_n$, 我们得到

$$\begin{aligned}\boldsymbol{B}_{21}\boldsymbol{A}_{11}+\boldsymbol{B}_{22}\boldsymbol{A}_{21}&=\boldsymbol{0}\\ \boldsymbol{B}_{21}\boldsymbol{A}_{12}+\boldsymbol{B}_{22}\boldsymbol{A}_{22}&=\boldsymbol{I}_q\end{aligned}$$

因为 $\boldsymbol{A}_{11}$ 非奇异, 存在 $\boldsymbol{A}_{11}^{-1}$. 解出 $\boldsymbol{B}_{11},\boldsymbol{B}_{12},\boldsymbol{B}_{21},\boldsymbol{B}_{22}$, 我们得到

$$\begin{aligned}\boldsymbol{B}_{11}&=\boldsymbol{A}_{11}^{-1}+(\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})\boldsymbol{D}^{-1}(\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1})\\ \boldsymbol{B}_{12}&=-(\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})\boldsymbol{D}^{-1}\\ \boldsymbol{B}_{21}&=-\boldsymbol{D}^{-1}(\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1})\\ \boldsymbol{B}_{22}&=\boldsymbol{D}^{-1}\end{aligned}$$

其中

$$\boldsymbol{D}=\boldsymbol{A}_{22}-\boldsymbol{A}_{21}(\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})$$

为了说明这些公式的使用, 对于分块矩阵

$$\boldsymbol{A}=\begin{pmatrix}1&2&3\\2&3&2\\3&3&4\end{pmatrix}$$

其中

$$\boldsymbol{A}_{11}=(1),\ \boldsymbol{A}_{12}=(2,3),\ \boldsymbol{A}_{21}=\begin{pmatrix}2\\3\end{pmatrix},\ \boldsymbol{A}_{22}=\begin{pmatrix}3&2\\3&4\end{pmatrix}$$

在这种分块情况下, $\boldsymbol{A}_{11}^{-1}=1$, 并且

$$\boldsymbol{D}=\begin{pmatrix}3&2\\3&4\end{pmatrix}-\begin{pmatrix}2\\3\end{pmatrix}(1)(2,3)=\begin{pmatrix}-1&-4\\-3&-5\end{pmatrix}$$

$$\boldsymbol{D}^{-1}=-\frac{1}{7}\begin{pmatrix}-5&4\\3&-1\end{pmatrix}=\begin{pmatrix}\frac{5}{7}&-\frac{4}{7}\\-\frac{3}{7}&\frac{1}{7}\end{pmatrix}$$

因此有

$$\boldsymbol{B}_{11}=\left(-\frac{6}{7}\right),\ \boldsymbol{B}_{12}=\begin{pmatrix}-\frac{1}{7}&\frac{5}{7}\end{pmatrix},\ \boldsymbol{B}_{21}=\begin{pmatrix}\frac{2}{7}\\\frac{3}{7}\end{pmatrix},\ \boldsymbol{B}_{22}=\begin{pmatrix}\frac{5}{7}&-\frac{4}{7}\\-\frac{3}{7}&\frac{1}{7}\end{pmatrix}$$

由此直接得到 $\boldsymbol{B}=\boldsymbol{A}^{-1}$.

D.2.8 用 Excel 进行矩阵运算

可以用 Excel 程序模块进行下列矩阵运算:

(1) 转置; (2) 乘法;

(3) 求非奇异矩阵的逆; (4) 求非奇异矩阵的行列式值.

图 D.1 是矩阵运算演示的例子 (文件 excelMatManip.xls). 在例 1(转置) 中, $\boldsymbol{A}$ 是一个 (2×3) 阶矩阵, 矩阵元素的输入范围是 A4:C5. 转置Transpose($\boldsymbol{A}$), 即 $\boldsymbol{A}^{\mathrm{T}}$, 出现在用户指定的 E4:F6 中, 得出该选定范围输出结果的计算步骤为:

Examples of Matrix Manipulations						
Example 1–Transpose (TRANSPOSE):						
Matrix A:				**Transpose(A):**		
10	20	30		10	100	
100	200	300		20	200	
				30	300	
Example 2–multiplication (MMULT):						
Matrix A:				**AB:**		
10	20	30		60		
11	21	31		63		
12	22	32		66		
13	23	33		69		
Matrix B:						
1						
1						
1						
Example 3--Inversion (MINVERSE):						
Matrix A:				**Inverse(A):**		
11	22	33		-0.20274	-0.01151	0.200837
16	-56	12		-0.03233	-0.01778	0.037657
17	19	34		0.119437	0.01569	-0.09205
Example 4--Determinant (MDETERM):						
Matrix A:				**Det(A):**		
1	2	3		239		
6	-4	5				
7	9	2				

图 D.1 用 Excel 作矩阵运算 (文件 excel Mat Manip.xls)

(1) 在 E4 单元格输入公式 =TRANSPOSE(A4:C5);

(2) 选择 (拉黑) 输出单元格范围 E4:F6;

(3) 按下 F2 键;

(4) 按下 CTRL+SHIFT+ENTER 键.

在例 2 中, 我们把输入矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的元素分别放在 A10:C13 和 A16:A18 中, 指定输出矩阵的范围 E10:E13. 下一步要在单元格 E10 输入公式 =MMULT(A10:C13, A16:A18), 并完全按照例 1 的第 2 步到第 4 步 (把 E4:F6 替换成 E10:E13). 注意到 MMULT(A16:A18, A10:C13) 没有定义. 在例 3 中, 单元格范围 A22:C24 内矩阵的逆通过在 E22 中输入公式 =MINVERSE(A22:C24) 被指定到 E22:G24 区域, 然后按照例 1 的第 2 步到第 4 步操作. 最后, 在例 4 中, 在单元格 E28 中输入公式 =MDETERM(A28:C30), 我们得到 A28:C30 内矩阵的行列式值.

D.3 二 次 型

设

$$\boldsymbol{X} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$$

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

则称函数

$$Q(X) = X^{\mathrm{T}} A X = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

为一个二次型. 我们总假定矩阵 A 是对称的, 因为每对系数 a_{ij} 和 $a_{ji}(i \neq j)$ 的每个元素都可以用 $\frac{a_{ij}+a_{ji}}{2}$ 代替而不改变 $Q(X)$ 的值.

举例说明. 二次型

$$Q(X) = (x_1, x_2, x_3) \begin{pmatrix} 1 & 0 & 1 \\ 2 & 7 & 6 \\ 3 & 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

等于

$$Q(X) = (x_1, x_2, x_3) \begin{pmatrix} 1 & 1 & 2 \\ 1 & 7 & 3 \\ 2 & 3 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

注意, 在第 2 种情况下, A 是对称矩阵. 因此以后总假定 A 是对称的.

称一个二次型是

(1) **正定的**, 若对所有的 $X \neq 0$, 有 $Q(X) > 0$.

(2) **半正定的**, 若对所有的 X, $Q(X) \geqslant 0$, 而且存在 $X \neq 0$, 使得 $Q(X) = 0$.

(3) **负定的**, 若 $-Q(X)$ 是正定的.

(4) **半负定的**, 若 $-Q(X)$ 是半正定的.

(5) **不定的**, 在所有其他情况下.

可以证明, 出现上述各情况的必要和充分条件如下.

(1) $Q(X)$ 是正定的 (半正定的), 若 A 的所有主子行列式的值都是正的 (非负的)[①]. 在这种情况下, 矩阵 A 称为是正定的 (半正定的).

(2) $Q(X)$ 是负定的, 若 A 的第 k 个主子行列式的值的符号是 $(-1)^k$, $k = 1, 2, \cdots, n$. 在这种情况下, 矩阵 A 称为是负定的 (半正定的).

(3) $Q(X)$ 是半负定的, 若 A 的第 k 个主子行列式要么为零, 要么符号是 $(-1)^k$, $k = 1, 2, \cdots, n$.

---

① $A_{n \times n}$ 的第 k 个主子行列式定义为

$$\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{vmatrix}, \quad k = 1, 2, \cdots, n.$$

D.4 凸函数和凹函数

函数 $f(\boldsymbol{X})$ 称为是一个严格凸函数, 若对于任意两个不同的点 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$, 都有

$$f(\lambda\boldsymbol{X}_1+(1-\lambda)\boldsymbol{X}_2)<\lambda f(\boldsymbol{X}_1)+(1-\lambda)f(\boldsymbol{X}_2)$$

其中 $0<\lambda<1$. 反过来, 一个函数 $f(\boldsymbol{X})$ 称为严格凹的, 若 $-f(\boldsymbol{X})$ 是严格凸的.

凸函数 (凹函数) 的一种特殊情况是二次型 (见附录 D.3)

$$f(\boldsymbol{X})=\boldsymbol{C}\boldsymbol{X}+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{X}$$

其中 $\boldsymbol{C}$ 是一个常数向量, $\boldsymbol{A}$ 为一个对称矩阵. 可以证明, 若 $\boldsymbol{A}$ 是正定矩阵, 则 $f(\boldsymbol{X})$ 是严格凸函数, 若 $\boldsymbol{A}$ 是负定矩阵, 则 $f(\boldsymbol{X})$ 是严格凹的.

习题

1. 说明下列向量是线性相关的:

(a) $\begin{pmatrix}1\\-2\\3\end{pmatrix}\begin{pmatrix}-2\\4\\-2\end{pmatrix}\begin{pmatrix}1\\-2\\-1\end{pmatrix}$ (b) $\begin{pmatrix}2\\-3\\4\\5\end{pmatrix}\begin{pmatrix}4\\-6\\8\\10\end{pmatrix}$

2. 已知

$$\boldsymbol{A}=\begin{pmatrix}1&4&9\\2&5&-8\\3&7&2\end{pmatrix},\quad \boldsymbol{B}=\begin{pmatrix}7&-1&2\\9&4&8\\3&6&10\end{pmatrix}$$

求:

(a) $\boldsymbol{A}+7\boldsymbol{B}$ (b) $2\boldsymbol{A}-3\boldsymbol{B}$ (c) $(\boldsymbol{A}+7\boldsymbol{B})^{\mathrm{T}}$

3. 在第 2 题中, 证明 $\boldsymbol{AB}\neq\boldsymbol{BA}$.

4. 考虑分块矩阵

$$\boldsymbol{A}=\begin{pmatrix}1&5&7\\2&-6&9\\3&7&2\\4&9&1\end{pmatrix},\quad \boldsymbol{B}=\begin{pmatrix}2&3&-4&5\\1&2&6&7\\3&1&0&9\end{pmatrix}$$

用分块矩阵法求 $\boldsymbol{AB}$.

5. 在第 2 题中, 用下列方法求 $\boldsymbol{A}^{-1}$ 和 $\boldsymbol{B}^{-1}$.

(a) 伴随矩阵法. (b) 行运算法.

(c) 逆矩阵乘积形式. (d) 分块矩阵法.

6. 考虑

$$\boldsymbol{B}=\begin{pmatrix}2&1&2\\0&2&1\\4&0&5\end{pmatrix},\quad \boldsymbol{B}^{-1}=\begin{pmatrix}\frac{5}{4}&-\frac{5}{8}&-\frac{3}{8}\\\frac{1}{2}&\frac{1}{4}&-\frac{1}{4}\\-1&\frac{1}{2}&\frac{1}{2}\end{pmatrix}$$

假定第 3 个向量 $\boldsymbol{P}_3$ 用 $\boldsymbol{V}_3 = \boldsymbol{P}_1 + 2\boldsymbol{P}_2$ 代替，这意味着所得到的矩阵是奇异的. 说明如何用逆矩阵乘积形式发现矩阵的奇异性.

7. 利用逆矩阵乘积形式检验下列哪组方程有唯一解、没有解、有无穷多个解.

(a) $x_1 + 2x_2 = 3$
$x_1 + 4x_2 = 2$

(b) $x_1 + 2x_2 = 5$
$-x_1 - 2x_2 = -5$

(c) $x_1 + 7x_2 + x_3 = 5$
$4x_1 + x_2 + 3x_3 = 8$
$x_1 + 3x_2 - 2x_3 = 3$

8. 验证附录 B.2.7 给出的求分块矩阵逆的公式.

9. 求下列矩阵的逆矩阵：

$$\boldsymbol{A} = \begin{pmatrix} 1 & \boldsymbol{G} \\ \boldsymbol{H} & \boldsymbol{B} \end{pmatrix}, \quad \boldsymbol{B}\text{非奇异}$$

10. 证明以下二次型是负定的：

$$Q(x_1, x_2) = 6x_1 + 3x_2 - 4x_1x_2 - 2x_1^2 - 3x_2^2 - \frac{27}{4}$$

11. 证明以下二次型是正定的：

$$Q(x_1, x_2, x_3) = 2x_1^2 + 2x_2^2 + 3x_3^2 + 2x_1x_2 + 2x_2x_3$$

12. 证明函数 $f(x) = \mathrm{e}^x$ 在整个实数轴上是严格凸函数.

13. 证明二次函数

$$f(x_1, x_2, x_3) = 5x_1^2 + 5x_2^2 + 4x_3^2 + 4x_1x_2 + 2x_2x_3$$

是严格凸的.

14. 在第 13 题中，证明 $-f(x_1, x_2, x_3)$ 是严格凹函数.

参 考 文 献

Hadley, G., *Matrix Algebra,* Addison-Wesley, Reading, MA, 1961.

Hohn, F., *Elementary Matrix Algebra,* 2nd ed., Macmillan, New York, 1964.

Press, W., B. Flannery, B. Teukolsky, and W. Vetterling, *Numerical Recipes: The Art of Scientific Computing,* Cambridge University Press, Cambridge, England, 1986.

附录 E　应 用 案 例

第 2 章应用题

2-1[①]Hi-V 公司生产和罐装 3 种橙汁产品：浓缩橙汁、普通橙汁和果酱. 这些产品出于大宗销售目的, 加工成 5 加仑的大桶. 果酱要用一级橙果, 其他两种产品用二级橙果. 表 E.1 列出了橙果的用途以及下一年的需求量. 市场调查表明, 对普通橙汁的需求量至少是浓缩橙汁需求量的两倍.

表　E.1

产品	橙果等级	每 5 加仑罐装产品所用的橙果磅数	最大需求量 (罐)
果酱	I	5	10 000
浓缩橙汁	II	30	12 000
普通橙汁	II	15	40 000

过去, Hi-V 公司分别按照每磅 25 美分和 20 美分的价格采购一级和二级橙果. 而今年, 一场突如其来的霜冻迫使果农提前收获和销售橙果, 分不出一级和二级来. 据估计, 300 万磅橙果中, 有 30% 属于一级果, 只有 60% 属于二级果. 为此, 所有的收成按照统一的销售价格出售, 每磅 19 美分. 按 Hi-V 公司的估算, 要把这些橙果分类成一级果和二级果, 公司要花费 2.15 美分/磅, 达不到标准橙果的 (占总收成的 10%) 将要扔掉.

为了成本分配的目的, 财务部门采用下面的依据来估算每磅一级果和二级果的成本费用. 因为采购的橙果中有 10% 低于二级果的标准, 每磅的有效平均成本按 $\frac{(19+2.15)}{0.9}=23.5$ 美分计算. 由于采购批量中一级果和二级果的比例是 1:2, 相应的平均每磅采购成本费用按照过去的价格就是 $\frac{(20\times2+25\times1)}{3}=21.67$ 美分. 因此, 平均价格的增加部分 (=23.5 美分 −21.67 美分) 应按照 1 : 2 的比例重新分配到两种级别的橙果上, 得到一级果的费用为每磅 $25+1.83(\frac{1}{3})=25.61$ 美分, 二级果的费用为每磅 $20+1.83(\frac{2}{3})=21.22$ 美分. 利用这些信息, 财务部门编制了一份这 3 种产品的收益表, 见表 E.2.

表　E.2

	产品 (5 加仑罐)		
	果　酱	浓缩果汁	普通橙汁
售价	\$15.50	\$30.25	\$20.75
可变成本	9.85	21.05	13.28
固定分摊成本	1.05	2.15	1.96
总成本	\$10.90	\$23.20	\$15.24
净利润	4.60	7.05	5.51

① 资料来源: "Red Brand Canners," *Stanford Business Cases 1965*, Graduate School of Business, Stanford University.

为 Hi-V 公司制定一个生产计划.

2-2[①]某钢铁公司有一个铸造厂和两个轧钢厂. 铸造厂浇铸 3 种钢锭, 钢锭在运送到轧钢厂之前要在机械车间进行加工, 加工好的钢锭由轧钢厂制造成各种产品.

每个季度初, 两个轧钢厂各自编制一份钢锭的月度需求计划, 交给铸造厂. 然后铸造厂经理再根据机械加工车间的最大生产能力编制一份生产计划, 不足的部分通过直接用高价从外面采购来满足. 表 E.3 是铸造厂自己生产每个钢锭的费用和外部采购价格之间的对比. 然而, 厂管理部门指出, 这种供应不足的情况并不常见, 估计只有约 5% 的时候才会发生.

表 E.3

钢锭类型	重量 (磅)	内部成本 (每只钢锭 $)	外部采购价格 (每只钢锭 $)
1	800	90	108
2	1 200	130	145
3	1 650	180	194

表 E.4 列出了机械加工车间 4 台不同机床上的加工时间, 两个轧钢厂未来 3 个月对钢锭的需求量见表 E.5.

表 E.4

机床类型	每只钢锭加工时间			机床数量	每月每台机床的工作小时
	钢锭 1	钢锭 2	钢锭 3		
1	1	5	7	10	320
2	0	4	6	8	310
3	6	3	0	9	300
4	3	6	9	5	310

表 E.5

月度	钢锭需求量					
	轧钢厂 1			轧钢厂 2		
	钢锭 1	钢锭 2	钢锭 3	钢锭 1	钢锭 2	钢锭 3
1	500	200	400	200	100	0
2	0	300	500	300	200	200
3	100	0	300	0	400	200

为该机械加工车间制定一个生产进度安排.

2-3 ArkTec 公司为个人客户组装 PC 机. 未来 4 个季度的订单分别为 400, 700, 500, 200 台. ArkTec 也可以一个季度内生产超过需求量的台数, 在这种情况下, 每台计算机每个季度会发生 $100 的储存费用. 从一个季度到下个季度增加产量需要雇佣更多的工人, 这会使得该季度每台计算机增加生产成本 $60. 此外, 从一个季度到下个季度降低产量就需要解雇工人, 也会使得该季度每台计算机的生产成本增加 $50.

① S. Jain, K. Stott, and E. Vasold, "Orderbook Balancing Using a Combination of Linear Programming and Heuristic Techniques," *Interfaces*, Vol.9, No.1, pp.55-67, 1978.

ArkTec 公司该如何安排计算机的组装线来满足这 4 个季度的需求呢?

2-4 Beaver 家具公司生产椅子、桌子和书架. 工厂先生产出半成品, 然后在公司的组装线上进行组装. 工厂 (组装前) 月度生产能力包括 3 000 把椅子、1 000 张桌子和 580 个书架. 组装线雇用了 150 个工人, 每天两班, 每班工作 8 小时, 每周 5 天工作日. 每把椅子、每张桌子和每个书架的平均组装时间分别为 20 分钟、40 分钟和 15 分钟.

因为工人每年有休假, 所以在组装线上工作的工人数量是浮动的. 请求休假的人数是, 5 月份有 20 人, 6 月份 25 人, 7 月份 45 人. 表 E.6 给出了市场部门预测的这 3 种产品 5 月、6 月和 7 月的销售量. 这 3 种产品的生产成本和销售价格见表 E.7. 如果生产出来的每件产品在某个月没有卖出, 则存放起来下个月销售, 储存费用大约是生产成本的 2%.

表 E.6

产品	预测销售量			4 月末库存量
	5 月份	6 月份	7 月份	
椅子	2 800	2 300	3 350	30
桌子	500	800	1 400	100
书架	320	300	600	50

表 E.7

产品	单位成本 ($)	销售单价 ($)
椅子	150	250
桌子	400	750
书架	60	120

Beaver 公司应该批准这些工人提出的休年假请求吗?

第 3 章应用题

3-1 一家小型罐头食品公司生产 5 种罐头食品, 这些产品用 3 种新鲜水果作为原料. 生产过程用到两个生产部门, 原来设计了富余的生产能力, 用于将来扩产. 事实上, 该公司现在按照一个班组生产, 随着需求量的增加, 很容易能扩产到两个班和 3 个班生产. 当前的实际限制条件是没有足够的新鲜水果. 由于公司自己冷库的能力有限, 必须每天采购鲜果.

公司刚刚聘用了一位年轻的学运筹学的分析人员, 对生产情况作了分析以后, 他决定为工厂建立一个 LP 模型. 该模型含有 5 个决策变量 (5 种产品), 3 个约束条件 (原材料). 利用这 3 个约束和 5 个变量, 按照 LP 的理论, 最优解是最多只能有 3 种产品. “啊哈,” 这位分析人员想, “这个公司的生产没达到最优.”

该分析人员去见了工厂经理, 详细讨论他的 LP 模型. 这位经理似乎弄明白了建模的概念, 同意这位分析人员的模型, 认为它接近于实际情况的表达. 这位分析人员接着解释说, 根据线性规划的理论, 最优的产品种数不应该超过 3 种, 因为这模型只有 3 个约束. 因为如此, 建议考虑让两种利润不好的产品停产.

经理仔细听取了他的汇报, 然后告诉这位分析人员说, 公司坚决要生产所有这 5 种产品, 因为出于市场竞争的考虑, 而且公司不可能停产任何一种产品. 这位运筹学分析人员回答道, 要想解决这种情况, 唯一的办法就是至少再添加两个约束, 这种情况下, 最优的 LP

模型才有可能包括所有这 5 种产品. 这时, 经理搞糊涂了, 因为要生产更多的产品就必须添加更多的约束条件的想法根本不符合最优性. 这位分析人员回答道:"这就是 LP 理论所说的呀!"

你对这种"自相矛盾"怎么解释呢?

3-2[①]一家 LTL 运输公司专门经营小件运输, 其运送点遍及全美国. 当货物到达某个运送点后, 要分拣这些货物, 有些送到当地客户, 有些再转运到其他地点. 运送点的货站雇用的工作人员有固定工和临时工. 固定工是一些工会雇员, 保证每周能工作 40 个小时, 一个固定工被分配到一天标准三班中的一班, 连续 5 天上同一个班, 但可能从一周的任何一天开始. 临时工是临时雇用的, 工作若干小时, 为可能超过现有固定工工作量的高峰时段提供帮手. 与工会签订的合同要求, 临时工每周的工作时间必须少于 40 小时.

运输的货物在一天内的任何时间都可能到达运送点, 为了实用目的, 我们假定运送的货物量随一天的时间连续变化. 对历史数据的分析表明, 货物量水平每周呈循环模式, 周末 (星期五到星期日) 达到高峰. 公司的政策要求: 一件货物必须在到达运送点后 16 个小时内处理完毕.

请建立一个模型, 求出固定工的每周工作安排方案.

3-3[②]Elk Hills 油田由美国联邦政府控股 (80%), 其能源部 (DOE) 按法律授权将所生产的政府股份的石油销售给最高竞价人. 同时, 法律也对向任何一个竞价单位供应的油量设定了上限. 该油田有 6 个供应点, 其生产能力 (bbl/天) 各不相同, 每个供应点的日产量 (bbl/天) 作为竞价项每天公布, 一个竞价人可以对任意多个竞价项提出竞争价格. DOE 接收这些竞价并对它们进行评估, 从竞价项 1 开始, 一直到竞价项 6, 向出最高价的竞价人供油, 但要考虑法律规定的能供应给一个竞价人的油量上限. 具体地, 表 E.8 给出了某一天红利价格的竞价情况. 红利指在供应点地区生产的同等级石油最高报价之上的加价. 任何竞价人购买的油量不得超过所有供应点日产总量 180 000 bbl 的 20%.

表 E.8

竞价项	竞价人出价的红利 ($/bbl)								产量
	1	2	3	4	5	6	7	8	(1 000 bbl/天)
1	1.10	0.99	1.20	1.10	0.95	1.00	1.05	1.02	20
2	1.05	1.02	1.12	1.08	1.09	1.06	1.11	1.07	30
3	1.00	0.95	0.97	0.94	0.93	1.01	1.02	0.98	25
4	1.30	1.25	1.31	1.27	1.28	1.26	1.32	1.32	40
5	1.09	1.12	1.15	1.07	1.08	1.11	1.05	1.10	35
6	0.89	0.87	0.90	0.86	0.85	0.91	0.88	0.91	30

DOE 采用一种评分排队的方法按照报价高低分配石油. 从竞价项 1 开始, 竞价人 3 给出的红利报价最高 (=$1.2), 因此他得到竞价项 1 以及法律要求的 20% 上限 ($= 0.2 \times 18\,000 = 36\,000$ bbl) 所允许的最大供油量. 根据表中的数据, 所有竞价项 1 的产量 (20 000

① 资料来源于作者对一家全国 LTL 运输公司开展的一项研究项目.

② 资料来源: B. Jackson and J. Brown, "Using LP for Crude Oil Sales at Elk Hills: A Case Study," *Interfaces*, Vol.10, pp.65-69, No.3, 1980.

bbl) 都分配给了竞价人 3. 接下来看竞价项 2, 还是竞价人 3 给出了最高的红利价, 但最多只能够供给他 16 000 bbl, 因为有 20% 的限制. 剩余的量给了第二高红利报价的竞价人 (= \$1.11), 因此给竞价人 7 分配 14 000 bbl(=30 000−16 000). 重复着一过程直到竞价项 6 分配完毕.

本问题所提出的方法能保证政府的最大日收益吗? 政府能通过提高或降低 20% 的限制取得更好的收益吗?

第 4 章应用题

4-1[①]MANCO 公司生产 P1, P2, P3 3 种产品, 生产过程用到原材料 R1 和 R2, 这些原材料由设备 F1 和 F2 加工. 表 E.9 给出了该问题的相关数据. 对 P2 的最小日需求量是 70 件, 对 P3 的最大需求量为 240 件. P1, P2, P3 的单位收益分别是 \$300, \$200, \$500.

表 E.9

资源	单位	每件产品用量			最大日产量
		P1	P2	P3	
F1	分	1	2	1	430
F2	分	3	0	2	460
R1	磅	1	4	0	420
R2	磅	1	1	1	300

MANCO 的管理部门正在研究改善公司财务状况的办法. 讨论下列方案的可行性:

(1) P3 的每单位收益可以增加 20%, 但这会让市场需求从 240 件减少到 210 件.

(2) 原材料 R2 在限制现有产量中似乎是一个关键因素. 按照比现有供应商的每磅高 \$3 的价格再寻找一家供应商, 则可增加产量.

(3) F1 和 F2 的加工能力每天可增加 40 分钟, 每台设备每天增加加工成本费 \$35.

(4) 产品 P2 的主要买家提出, 每天的供应量可从现在的 70 件提高到 100 件.

(5) P1 在 F2 上的每件加工时间可从 3 分钟减少到 2 分钟, 每天增加成本 \$4.

4-2 Reddy Mikks 公司正在制定一项未来发展计划. 一份市场调查表明, 公司可增加销售量 25%. 为了制订一个行动计划, 公司正在对下列方案进行研究 (详细的模型和求解见例 3.3-1).

方案 1 因为增加 25% 的销售量相当于增加了约 \$5 250 的收入, 并且每增加 1 吨 M1 和 M2 分别增加价值 \$750 和 \$500, 要实现所要求的提高产量, 就需要每种 M1 和 M2 增加 $\$5250 \div \frac{(\$750+\$500)}{2} = 8.4$ 吨.

方案 2 原材料 M1 和 M2 的增加量分别为 6 吨和 1 吨. 这些增量等于 M1 和 M2 现有水平 (分别等于 24 吨和 6 吨) 的 25%. 由于这两种资源在当前的最优解中是短缺的, 增加 25% 的供货量就能得到同样水平的所需要的内外涂料的产量增长.

你对这些方案的看法如何? 你能提出解决这个问题的其他方法吗?

---

① 资料来源: D. Sheran, "Post-Optimal Analysis in Linear Programming–The Right Example," *IIE Transactions*, Vol.16, No.1, pp.99-102, March 1984.

第 5 章应用题

5-1[①]ABC 可乐公司在 Tawanda 岛国北部设了一家工厂, 生产 3 种包装的软饮料, 包括可回收型玻璃瓶、铝罐和不可回收的塑料瓶. 可回收的 (空) 瓶运送到销售库房, 在工厂内再利用.

因为需求量的连续增长, ABC 公司想要另外建一家工厂. 表 E.10 给出了未来 5 年对软饮料的需求情况 (单位: 箱), 表 E.11 是现有工厂这 5 年期间的规划生产能力. 该公司拥有 6 个分销仓库: N1 和 N2 在北部、C1,C2 在中部, S1 和 S2 位于南部. 表 E.12 给出了每个仓库在本地区的销售份额. 接近于 60% 的销售量在北部, 15% 在南部, 而 25% 的销售量出现在南部地区.

表 E.10

包装方式	年度				
	1	2	3	4	5
可回收的玻璃瓶	2 400	2 450	2 600	2 800	3 100
易拉罐	1 750	2 000	2 300	2 650	3 050
不可回收的塑料瓶	490	550	600	650	720

表 E.11

包装方式	年度				
	1	2	3	4	5
可回收的玻璃瓶	1 800	1 400	1 900	2 050	2 150
易拉罐	1 250	1 350	1 400	1 500	1 800
不可回收的塑料瓶	350	380	400	400	450

表 E.12

仓库	销售份额百分比
N1	85
N2	15
C1	60
C2	40
S1	80
S2	20

该公司想要在中部或者南部建设一家新厂. 每箱可回收的瓶子的运输成本见表 E.13. 按照估算, 每箱易拉罐和每箱不可回收的瓶子的运输成本分别是可回收的瓶子的 60% 和 70%.

表 E.13

仓库	每箱运输成本 ($)		
	现有厂	中部厂	南部厂
N1	0.80	1.30	1.90
N2	1.20	1.90	2.90
C1	1.50	1.05	1.20
C2	1.60	0.80	1.60
S1	1.90	1.50	0.90
S2	2.10	1.70	0.80

① 资料来源: T. Cheng and C. Chiu, "A Case Study of Production Expansion Planning in a Soft-Drink Manufacturing Company," *Omega,* Vol.16, No.6, pp.521-532, 1988.

新厂的位置应该设在该国的中部还是南部?

5-2[①]建设布里斯班国际机场需要将从附近海湾的 5 个沙场挖掘出来的约 1 355 000 立方米的沙子用管道运送到机场工地的 9 处施工现场. 这些沙子用作稳固施工区域的潮湿地面. 一些要送来沙子的现场专门修建了场内和机场周围的道路. 对于现场中的多余的沙子, 将用卡车运到机场周边将要修建道路的其他偏远地区. 沙场到施工现场的距离 (单位: 100 米) 见表 E.14, 这个表还说明了每个施工现场地点的供应量和需求量.

表 E.14

	1	2	3	4	5	6	7	8	9	供应量
1	22	26	12	10	18	18	11	8.5	20	**960**
2	20	28	14	12	20	20	13	10	22	**201**
3	16	20	26	20	1.5	28	6	22	18	**71**
4	20	22	26	22	6	∞	2	21	18	**24**
5	22	26	10	4	16	∞	24	14	21	**99**
需求量	**62**	**217**	**444**	**315**	**50**	**7**	**20**	**90**	**150**	

(a) 该项目的管理部门已经估算过, 要运送 2 495 000 单位 [体积 (m^3)× 距离 (100 m)] 沙子, 每单位费用为 \$0.65. 项目管理部门给出的这个运送沙子的估算准确吗?

(b) 项目管理部门已经认识到, 必须要在道路建设完成后, 才能把沙子运送到施工现场. 尤其是, 外围道路 (目的地 9) 必须在向某些现场运送沙子之前建成. 在表 E.15 中, 用 x 标出了不通道路的情况, 它们需要完成外围道路的建设. 根据这些限制条件, 应如何运送沙子呢?

表 E.15

	1	2	3	4	5	6	7	8	9
1	x	x			x				
2	x	x			x				
3			x			x			
4			x					x	
5	x	x			x		x		

5-3 10 年前, 有个批发商开办了一家公司, 从中心货栈 (CW) 销售药品. 订单药品用面包车送到顾客那里, 随着需求的增长, 他扩大了中心货栈, 此外, 还新建了 W1 和 W2 两个新货栈. 中心货栈的存货通常比较全, 偶尔也会为新建货栈供应某些缺货. 现在, 这种偶尔缺货式的供应已经发展成一种大规模的经营方式, 这两个新建货栈有 $\frac{1}{3}$ 的分销库存来自中心货栈. 表 E.16 给出了这 3 个货栈向顾客点 C1 到 C6 发送的订单量. 一个顾客点是一个城镇, 设有若干药店.

① 资料来源: C. Perry and M. Ilief, "Earth Moving on Construction Projects," *Interfaces*, Vol.13, No.1, pp.79-84, 1983.

表 E.16

线路		订单数	线路		订单数
从	到		从	到	
CW	W1	2 000	W1	C3	1 100
CW	W2	1 500	W1	C4	1 500
CW	C1	4 800	W1	C5	1 800
CW	C2	3 000	W2	C2	1 900
CW	C3	1 200	W2	C5	600
W1	C1	1 000	W2	C6	2 200

几年来, 批发商的供货安排逐渐发展到现在的状态. 其实, 这种供货安排过去是按照某种分散式的方式, 由每个货栈根据"自主经营"的标准来决定供货区域. 的确, 在某些情况下, 货栈经理们会争夺新的客户, 主要为了提高他们的"影响范围". 例如, 中心货栈的经理们声称, 他们的供货区域不仅包括普通的顾客, 还有其他两个货栈. 在很多情况下, 同一个城镇 (顾客点) 内有好几家货栈向不同的药店供应药品.

不同的顾客点之间面包车的行车距离 (英里) 见表 E.17. 一车通常可装 100 个订单的货物量. 请对该批发商现在的销售策略做出评价.

表 E.17

	CW	W1	W2	C1	C2	C3	C4	C5	C6
CW	0	5	45	50	30	30	60	75	80
W1	5	0	80	38	70	30	8	10	60
W2	45	80	0	85	35	60	55	7	90
C1	50	38	85	0	20	40	25	30	70
C2	30	70	35	20	0	40	90	15	10
C3	30	30	60	40	40	0	10	6	90
C4	60	8	55	25	90	10	0	80	40
C5	75	10	7	30	15	6	80	0	15
C6	80	60	90	70	10	90	40	15	0

5-4 Kee Wee 航空公司经营 Waco 与 Macon 之间的 8 个往返航班, 航班时刻表见表 E.18. 如果在另一个城市至少有 90 分钟的停留时间, 机组人员就能够在当天返回到其本地机场 (Waco 或 Macon); 否则, 机组只能第二天返回. 要求将机组与从这两个城市出发的航班之间做出编排, 使得所有机组的总停留时间最短.

表 E.18

航班	从 Waco 出发	到达 Macon	航班	从 Waco 出发	到达 Macon
W1	6:00	8:30	M1	7:30	9:30
W2	8:15	10:45	M2	9:15	11:15
W3	13:30	16:00	M3	16:30	18:30
W4	15:00	17:30	M3	20:00	22:00

第 6 章应用题

6-1 一位住在旧金山 (SF) 的户外运动爱好者想利用一次 15 天的假期去 4 个国家公园：优山美地 (YO), 黄石 (YE), 大提顿 (GT), 罗斯摩尔山 (MR). 这次旅行从旧金山出发, 再回到旧金山, 所去公园的路线是 SF→YO→YE→ GT→MR→SF, 并且在每个公园呆上两天时间. 从一个公园到下一个公园, 要么乘飞机, 要么开车. 如果坐飞机, 每段旅程需要 1/2 天时间; 如果开车的话, 从 SF 到 YO 需要 1/2 小时, 从 YO 到 YE 需要 3 天, 从 YE 到 GT 要 1 天, 从 GT 到 MR 需要 2 天, 而从 MR 回到 SF 则需要 3 天. 乘坐汽车的好处是便宜, 但需要的时间长. 考虑到这个人要在 15 天后必须回来上班, 目标就是要在 15 天的期限内让旅行的费用最少. 表 E.19 给出了乘坐飞机和开车旅行的单程费用, 求旅行中每个旅段的最优旅行方式.

表 E.19

出发地	乘飞机旅行的费用 ($)					乘汽车旅行的费用 ($)				
	SF	YO	YE	GT	MR	SF	YO	YE	GT	MR
SF	–	150	350	380	450	–	130	175	200	230
YO	150	–	400	290	340	130	–	200	145	180
YE	350	400	–	150	320	175	200	–	70	150
GT	380	290	150	–	300	200	145	70	–	100
MR	450	340	320	300	–	230	180	150	100	–

6-2[①] 有个捐赠人要给 Springdale 公共图书馆捐赠图书, 这些书有 4 种高度：12 英寸、10 英寸、8 英寸和 6 英寸. 馆长估计, 对 12 英寸的书籍需要 12 英尺的书架, 10 英寸的书籍需要 18 英尺的书架, 8 英寸的书籍需要 9 英尺的书架, 而 6 英寸的书籍则需要 10 英尺的书架. 制作一个书架的费用包括每英尺长度的固定费用和可变费用, 如表 E.20 所示.

表 E.20

书架高度 (ft)	固定费用 ($)	可变费用 ($/ft)
12	25	5.50
10	25	4.50
8	22	3.50
6	22	2.50

假设允许较小的书籍放在较大的书架上, 那么该如何设计这些书架呢?

6-3 一家船运公司想要把 5 船货物从 A, B, C 港口运送到 D 和 E 港口. 这 5 次运输的交货日期如表 E.21 所示. 表 E.22 给出了这些港口之间 (假设往返航行时间要短一些) 的行船时间 (天). 该公司想要知道按照给定的船运计划所需要的最少船数是多少.

① 资料来源：A. Ravindran, "On Compact Storage in Libraries," *Opsearch*, Vol.8, No.3, pp.245-252, 1971.

表 E.21

送货	行船路线	到货日期
1	A 到 D	10
2	A 到 E	15
3	B 到 D	4
4	B 到 E	5
5	C 到 E	18

表 E.22

	A	B	C	D	E
A				3	4
B				3	2
C				3	5
D	2	2	2		
E	3	1	4		

6-4[①]有几个人曾经一起在海外设立了一家自由经纪人公司, 经营高投机性股票. 这些经纪人在一种允许各种经纪人之间交易的宽松金融系统下运作, 包括买入、卖出、借贷和放贷. 对于整个这一批经纪人来说, 他们的收入来自对外客户销售所收取的佣金.

最终, 投机股票中的风险交易变得一塌糊涂, 所有的经纪人都破了产. 那时的财务状况是, 所有的经纪人都欠外部客户的债, 而且经纪人之间的金融纠缠极其复杂, 几乎每一个经纪人都欠公司内其他经纪人的债.

那些个资产能够支付欠债的经纪人被宣布有偿还能力. 剩下的经纪人求助于法律机构, 让他们从外部客户的利益出发来解决好债务问题. 由于没有偿还能力的经纪人的实际资产少于应支付量, 所有的债务按比例进行了分配. 最终的结果是, 将没有偿还能力的经纪人的所有资产都进行了完全清算.

在处理没有偿还能力的经纪人的内部财务纠缠问题时, 人们决定交易只是为了满足法律要求, 因为实际上, 没有一个经纪人手上还有别人的资金. 法律机构要求, 将经纪人之间的交易数减少到绝对最小数. 这就意味着, 若 A 欠 B X 元, 而且 B 欠 A Y 元, 这两笔"循环"交易就减少成一笔, 欠款额为 $|X-Y|$. 若 $X>Y$, 这笔钱就由 A 支付给 B, 反之若 $Y>X$, 则由 B 支付给 A, 假如是 $X=Y$, 则这两笔交易就完全抵消了. 这一思路被广泛用在解决涉及任意多个经纪人的循环债务问题.

你会怎样来处理这种情况呢? 具体地, 要求你回答以下两个问题.

(1) 应如何分摊债务?

(2) 应如何把经纪人之间的交易数减到最小?

第 7 章应用题

7-1[②]Warehouzer 公司经营 3 处木材生产林地, 植树面积分别是 100 000, 180 000, 200 000 英亩. 主要的木材产品包括 3 类, 分别是纸浆用木材、夹板和原锯木. 在每块林地上, 可以有几种植树方案, 每种方案的成本、轮种年限 (即从树苗到成材所需的年数)、土地租用收入以及产量都不一样. 表 E.23 是这些数据的汇总表.

为了保证未来生产的持续性, 每种方案中, 所种植的每英亩林地都要求, 英亩数等于方案制订的轮种年限数. 租金列表示每英亩的土地租用收入.

① 资料来源: H. Taha, "Operations Research Analysis of a Stock Market Problem," *Computers and Operations Research*, Vol.18, No.7, pp.597-602, 1991.

② 资料来源: K. Rustagi, *Forest Management Planning for Timber Production: A Goal Programming Approach*, Bulletin No.89, Yale University, New Haven, 1976.

表　E.23

林地	方案	年 $/英亩		轮种年限	年 m³/英亩		
		成本	租金		纸浆木	夹板	原木
1	A1	1 000	160	20	12	0	0
	A2	800	117	25	10	0	0
	A3	1 500	140	40	5	6	0
	A4	1 200	195	15	4	7	0
	A5	1 300	182	40	3	0	7
	A6	1 200	180	40	2	0	6
	A7	1 500	135	50	3	0	5
2	A1	1000	102	20	9	0	0
	A2	800	55	25	8	0	0
	A3	1 500	95	40	2	5	0
	A4	1 200	120	15	3	4	0
	A5	1 300	100	40	2	0	5
	A6	1 200	90	40	2	0	4
3	A1	1 000	60	20	7	0	0
	A2	800	48	25	6	4	0
	A3	1 500	60	40	2	0	4
	A4	1 200	65	15	2	0	3
	A5	1 300	35	40	1	0	5

Warehouzer 公司的目标是:

(1) 纸浆木、夹板和原木的年产量分别达到 200 000, 150 000 和 350 000 立方米.

(2) 年度植树预算是 250 万美元.

(3) 土地租金年收入达到每英亩 $100.

对于每块林地, 每种方案应该规划多大面积呢?

7-2 某个慈善机构开办了一家儿童福利院. 该机构利用义工为福利院提供服务, 服务时间每天从上午 8 点到下午 2 点. 志愿者可能从上午 8 点到 11 点之间的某个小时开始上班, 最多工作 6 小时, 最少 2 小时. 中午 12 点到下午 1 点是午饭时间, 这段时间大家都休息. 按照该慈善机构的估算, 全天 (从上午 8 点到下午 2 点, 除 12 点到下午 1 点午休时间外) 需要招募的义工目标数分别为 15, 16, 18, 20, 16 名, 目的是确定每个小时 (8:00, 9:00, 10:00, 11:00, 1:00) 开始上班的目标义工数, 使得尽可能满足需求. 请用目标规划模型表达并求解这个问题.

第 8 章应用题

8-1 一家开发公司在某个大城市市区拥有一块 90 英亩的土地, 想要建几幢办公大楼和一个商业中心. 建好的大楼经过 7 年出租, 以后再卖掉. 每幢大楼的售价预计是最后一年出租收入的 10 倍. 该公司估计, 该开发项目要建造一个 450 万平方英尺的商业中心. 按照项目计划, 需要建造 3 座高层写字楼和 4 座花园式办公楼.

该公司需要解决一个进度安排问题. 如果某幢大楼完工得太早, 就有可能空闲; 假如太迟, 则潜在的住户可能就去别的项目了. 根据市场调查, 未来 7 年对办公用房的需求情况

如表 E.24 所示, 而表 E.25 则列出了这 7 幢大楼的建筑面积.

表 E.24

需求 (1000 平方英尺)			需求 (1000 平方英尺)		
年	高层写字楼	花园办公楼	年	高层写字楼	花园办公楼
1	200	100	5	293	146
2	220	110	6	322	161
3	242	121	7	354	177
4	266	133			

表 E.25

花园楼	面积 (ft^2)	高层楼	面积 (ft^2)
1	60 000	1	350 000
2	60 000	2	450 000
3	75 000	3	350 000
4	75 000	–	–

办公楼租金毛收入预计在每平方英尺 \$25, 花园办公楼和高层办公楼的运行费分别为每平方英尺 \$5.75 和 \$9.75, 相应的建造施工成本分别为每平方英尺 \$70 和 \$105, 施工成本和租金收入预计每年增长 4%.

该公司应如何安排这 7 栋建筑的施工进度呢?

8-2[①]在一次全国高校体育协会的女子体操团体赛中, 设立了 4 个竞赛项目, 有跳马、高低杠、平衡木和自由体操. 每项比赛中, 每队可派 6 名运动员参加, 每个运动员的成绩按 1~10 分评估. A 大学队以往竞赛成绩如表 E.26 所示.

表 E.26

比赛项目	A 大学体操队运动员的成绩					
	1	2	3	4	5	6
跳马	6	9	8	8	4	10
高低杠	7	9	7	8	9	5
平衡木	9	8	10	9	9	8
自由体操	6	6	5	9	10	9

一个代表队的总成绩按照每个竞赛项目前 5 名成绩相加得出. 一名队员选择一种参赛方式, 要么参加单项比赛, 要么参加“全能”4 项比赛. 单项赛运动员最多参加 3 项比赛, 一个代表队至少必须派出 4 名队员参加全能赛. 请建立一个 ILP 模型, 用来选择竞赛队, 并求出最优解.

8-3[②]1990 年, 全美国大约有 180 000 家电话推销中心, 从业人员达 200 万人. 到 2000 年, 共

① 资料来源: P. Ellis and R. Corn, “Using Bivalent Integer Programming to Select Teams for Intercollegiate Women's Gymnastic Competition, ” *Interfaces*, Vol.14, No.3, pp.41-46, 1984.

② 资料来源: T. Spencer,A.Brigandi, D. Dargon, and M. Sheehan, “AT & T's Telemarketing Site Selection System Offers Customer Support,” *Interfaces*, Vol.20, No.1, pp.83-96, 1990.

有 700 000 多家公司雇佣了大约 800 万员工通过电话推销它们的产品, 因此到底要设立多少电话销售中心以及把它们安排在哪里就成了一个非常重要的问题.

ABC 公司正在考虑设立电话销售中心的数量以及地点. 公司可以考虑在几个候选地区选择设立一个中心, 可以为一个或几个地理区域提供 (部分或全部) 服务, 一个地理区域通常指一个或几个 (电话) 区号. ABC 公司的电话销售集中在 8 个区号: 501, 918, 316, 417, 314, 816, 502, 606. 表 E.27 给出了这些候选地点、它们的服务地区, 以及建立电话销售中心的费用. 表 E.28 是中心与不同区号之间每小时的通话费用.

表 E.27

中心地点	服务的区号	费用 ($)
得克萨斯州达拉斯市	501,918,316,417	500 000
佐治亚州亚特兰大市	314,816,502,606	800 000
肯塔基州路易维尔市	918,316,417,314,816	400 000
科罗拉多州丹佛市	501,502,606	900 000
阿肯色州小石城市	417,314,816,502	300 000
田纳西州孟菲斯市	606,501,316,417	450 000
密苏里州圣路易斯市	816,502,606,314	550 000

表 E.28

到 从	区号							
	501	918	316	417	314	816	502	606
得克萨斯州达拉斯市	$14	$35	$29	$32	$25	$13	$14	$20
佐治亚州亚特兰大市	$18	$18	$22	$18	$26	$23	$12	$15
肯塔基州路易维尔市	$22	$25	$12	$19	$30	$17	$26	$25
科罗拉多州丹佛市	$24	$30	$19	$14	$12	$16	$18	$30
阿肯色州小石城市	$19	$20	$23	$16	$23	$11	$28	$12
田纳西州孟菲斯市	$23	$21	$17	$21	$20	$23	$20	$10
密苏里州圣路易斯市	$17	$18	$12	$10	$19	$22	$16	$22

ABC 公司想要选定 3 到 4 个中心, 那么应该放在哪里呢?

8-4[①] 一家服务于农村地区的电力公司要确定客户服务的维修接线中心的数量和地点, 以便提供及时的维修和接线服务. 该公司按照居民居住地把顾客分成了 5 个群组, 如表 E.29. 公司已经选择了 5 个可能的维修接线中心地点, 表 E.30 汇总了从这些中心到不同居民点的平均距离. 服务卡车的平均行驶速度约为每小时 45 英里.

表 E.29

居民点	1	2	3	4	5
顾客人数	400	500	300	600	700

① 资料来源: T. Erkut, Myrdon, and K. Strangway, "Transatlanta Redesigns its Service Delivery Network," *Interfaces*, Vol.30, No.2, pp.54-69, 2000.

表 E.30

居民点	维修接线服务中心				
	1	2	3	4	5
1	40	100	20	50	30
2	120	90	80	30	70
3	40	50	90	80	40
4	80	70	110	60	120
5	90	100	40	110	90

如果公司希望把对顾客请求服务的响应时间限制在 90 分钟之内, 应建多少个维修接线中心呢?

8-5[①]在汽车制造业, 需要利用样车对一些新的设计方案进行测试. 制造这些样车的投资很大, 每台样车的制造费用可能会超过 $250 000. 测试的种类有很多, 由不同的部门所做的每一种测试都针对新设计方案的某些特殊性能来进行. 例如, 一台运输车型的性能可能包括车身形状、发动机大小、车顶高度、变速箱种类、后舱、整车重量以及底盘等. 为了检测最恶劣情况下高海拔地区的驾驶性能, 要求所制造样车的整车重量最大, 并配备自动的变速箱, 以及要求发动机最小. 车顶高度和底盘对于这类测试就不重要了.

开始时, 我们可以制造出一批样车, 满足测试部门指定的某些性能. 例如, 若测试 1 需要有性能 A, 测试 2 要求性能 B, 则我们可以造出两台不同的样车, 一台用于测试性能 A, 另一台测试性能 B. 还有一种方法是, 我们用两台完全一样的样车 (A, B) 来做这两项测试. 这样的好处是, 两台相同样车的造价比造两台不同样车的费用要便宜得多. 在这种情况下, (A, B) 被称作是*共用* 样车.

对于一般的情况, 令 $A_i, i = 1, 2, \cdots, n$ 为性能 i 的一组配置. 例如, 性能 1 是*变速箱*, 则 A_1={标准, 自动}. 用每种性能的一种配置造出一台样车. 这样, 如果性能 i 的配置数为 m_i 的话, 则最大可能的配置数就是 $\prod_{i=1}^{n} m_i$. 因此一套新的设计方案就可能涉及上千台样车, 我们的思路是, 对*共用*样车做出明智的选择, 使得能够满足测试部门的配置要求. 令 $B_j, j = 1, 2, \cdots, t$ 代表测试部门 j 要求的一组配置. 例如, B_1 ={V8 发动机, 标准发动机}. 因此集合 B 的元素必须是一台可以造得出的样车, 或者是某个子集.

根据以上信息, 该如何选择可以制造的样车, 以满足这些测试部门的要求. 运用已有的模型解决表 E.31 给出的问题. 表 E.32 是相关测试部门的要求. [注意到, 所给定的情形过于简单, 显然可以通过直接检验来求解. 实际情况下可能有上千个可制造的样车, 进行几十项测试内容, 求解这样的问题就不会那么简单了.]

表 E.31

性　能	配　置
发动机	{4 缸, 6 缸, 8 缸}
变速箱	{自动变速箱, 标准变速箱}
车身形状	{4 门, 双座, 敞篷}

① 资料来源: K. Chelst, J. Sidelko, A. Przebienda, J. Lockledge, and D. Mihailidis, "Rightsizing and Management of Prototype Vehicle Testing at Ford Motor Company," *Interfaces*, Vol.31, No.1, pp.91-107, 2001.

表 E.32

测试部门	要求的配置	测试部门	要求的配置
1	{4 缸, 标准变速箱}	4	{8 缸}
2	{8 缸, 双座}	5	{标准变速箱, 敞篷}
3	{6 缸, 敞篷}	6	{6 缸, 自动变速箱}

8-6[①]美国快运航空公司 (American Express Airlines) 经营 8 个城市 (C1 到 C8) 间的 18 个航班 (F1 到 F18), 有 10 个飞行机组 (R1 到 R10). 机组人员通常从一个指定的基地出发, 在完成指定的飞行任务后, 再返回到该基地. 表 E.33 是该航空公司每天的航班安排表. 公司的调度部门负责编制机组的飞行计划, 编制过程中需要考虑到各航段以及机组人员的要求. 一个可行的飞行路线由飞行员能够在计划期间提供飞行服务的若干航段组成, 表 E.34 给出了这 10 个机组可行的飞行路线. 把一个机组指派给一个飞行路线的费用与这条飞行路线所包括的航段数成比例.

表 E.33

出发地	目的地	航班号	出发地	目的地	航班号
C1	C2	F1	C3	C6	F10
C1	C3	F2	C4	C1	F11
C1	C5	F3	C4	C8	F12
C1	C8	F4	C5	C2	F13
C2	C4	F5	C5	C7	F14
C2	C7	F6	C6	C4	F15
C2	C8	F7	C6	C1	F16
C3	C1	F8	C7	C4	F17
C3	C2	F9	C8	C3	F18

表 E.34

机组	可行路线
1	(C3,C6,C4,C8,C3),(C3,C2,C8,C3)
2	(C1,C5,C7,C4,C1)
3	(C4,C8,C3,C2,C4),(C4,C1,C5,C7,C4)
4	(C1,C8,C3,C6,C4,C1),(C1,C5,C2,C1)
5	(C2,C4,C8,C3,C2),(C2,C4,C8,C3,C1,C2), (C2,C7,C4,C1,C2)
6	(C8,C3,C1,C8)
7	(C5,C2,C8,C3,C1,C5),(C5,C7,C4,C8,C3,C1,C5)
8	(C6,C1,C3,C6),(C6,C4,C1,C5,C7,C4,C8,C3,C6),(C6,C4,C8,C3,C6)
9	(C7,C4,C2,C7),(C7,C4,C8,C3,C2,C7)
10	(C1,C3,C6,C1),(C1,C2,C8,C3,C1),(C1,C8,C3,C1)

由于编制这些飞行路线必须要满足机组人员的要求, 同时要符合航空管理局 (FAA) 的规定, 由调度部门所提出的飞行路线不一定能覆盖所有航班, 也不一定满足所有机组人员

① **资料来源**: G. Yu, M. Arguello, G. Song, S. McCowan, and A. White, "A New Era for Recovery at Continental Airlines," *Interfaces*, Vol.3, No.1, pp.5-22, 2003.

的要求, 因此可能不是一个可行解. 我们的目标就是要在尽可能去掉不可行性的情况下指派机组的飞行路线. 我们假定, 若所得到的解是不可行的, 则调度部门应要么提出其他的可行飞行路线, 要么动用备份机组来完成飞行任务.

请建立一个模型, 用来评价计划部门提出的这些飞行路线, 并对解做出解释.

8-7[①]微电子器件的核心生产流程从制造**晶片**(如光盘一样的圆形硅薄片) 开始, 要把成千上万个集成电路蚀刻在晶片上. 做好的晶片再切割成小的矩形器件, 称为**部件**, 把它们放在一个支承材料上, 包好后就成为一个模块. 经过测试后, 这些电子部件被分成不同的类别, 每类具有不同的电路. 假定 N 为从一个有 n 类电路的晶片上切割出来的部件数, 归在 j 类中的部件数为 r_jN, 其中 $r_1+r_2+\cdots+r_n=1$, 对所有的 j 有 $r_j\geqslant 0$. 生产出的部件在制造模块时可以相互替换使用, 如一片部件 i 或者部件 j 可以用来生产一个模块 k. 给定分类的比率 r_j, 该生产出多少晶片才能满足这些模块的具体要求呢? 应如何把这些生产出来的部件分配到这些模块上去呢?

请利用表 E.35 和表 E.36 中的数据, 针对 5 个部件和 3 个模块的具体情况, 检验所建立的模型.

表 E.35

部件	1	2	3	4	5
每类部件的比例	0.21	0.19	0.1	0.3	0.2
初始库存	10	4	8	0	3

表 E.36

模块	需求 (个)	可互换部件
1	20	2, 3, 5
2	30	1, 3, 5
3	45	4 或 5

8-8[②]过去, 银行往往按照支票到达的随机顺序从顾客账户里兑现这些支票. 而现如今, 由于采用了现代数据处理系统, 允许一些银行合法地安排每天的支票支付顺序, 使得银行能够对资金不足的账户收取更高的退票费. 例如, 假设一个银行账户有 \$1 000 的余额, 而某一天连续收到了 4 张支票, 支票数额分别是 \$100,\$100,\$100, \$1 000, 如果这些支票按照收到的顺序进行支付, 只有一张支票 (\$1 000) 由于资金不足 (No-sufficient funds, NFS) 应该退回. 在这种情况下, 这位顾客要负担一笔 NSF 费用 (大约 \$20). 但是, 假如银行按照 \$1 000, \$100, \$100, \$100 的顺序来支付这些支票的话, 就会出现 3 笔 NSF, 银行就能收取 3 笔 NSF 费. 从这个例子可以看出, 好像银行只要通过从高到低的顺序, 先支付较大面

① 资料来源: P. Lyon, R. Milne, R. Orzell, and R. Rice, "Matching Assets with Demand in Supply-Chain Management at IBM Microelectronics," *Interfaces*, Vol. 31, No. 1, pp.108-124, 2001.

② 资料来源: A. Apte, U. Apte, R. Beatty, I. Sarkar, and J. Semple, "The Impact of Check Sequencing on NSF (Not-Sufficient Funds)Fees," *Interfaces*, Vol. 34, No. 2, pp. 97-105, 2004.

额的支票, 就能使得收取的 NSF 费达到最高, 但一般来讲不是这样的. 比如, 考虑一个余额为 \$1 200 的账户, 要从高到低支付 \$900, \$675, \$525, \$200, \$100, \$75, \$25 共 7 张支票. 在这种情况下, 可以兑现 \$900, \$200, \$100 这 3 张支票, 剩下的 4 张支票就要收取退票费. 而实际上, 如果银行不兑现这张 \$900 的支票, 而只是支付 \$675 和 \$525 两张支票 (=\$1 200), 它还可以多收一笔 NSF 费.

(a) 建立一个银行处理每日支票的模型, 保证银行收取的 NSF 费最高, 并利用上述给定的数据计算这个模型.

(b) 从道理上讲, 每个人都希望银行通过收取最低的 NSF 费为顾客提供最好的服务. 在这种情况下, 这些支票该怎样处理呢?

8-9 考虑应用题 3-3(第 3 章). 对某些竞价人, 只有满足表 E.37 中给定的最小需求量时才能得到供油. 一个成功的竞价人必须至少得到最小需求 (但在法律规定的 20% 限制之内), 否则得不到供油. 如何才能实现这项任务呢?

表 E.37

竞价项	1~6 竞价项 (1 000 桶) 的竞价人 1~8 要求的最小量							
	1	2	3	4	5	6	7	8
1	10	13	10	25	18	14	20	14
2	14	11	11	20	17	16	18	14
3	20	15	20	10	16	16	17	17
4	11	17	15	6	20	16	17	8
5	15	29	18	12	14	10	15	10
6	18	20	19	22	18	8	19	15

8-10[①] 一家建筑公司获得了 8 个项目的施工合同. 这些项目分布在美国的不同地区. 公司有 5 名项目经理, 每人负责一个项目. 这些经理们来自全国各地, 去往项目工地的旅行时间各不相同. 费用大的项目管理上要求得更多. 为了平衡起见, 公司对经理负责项目的安排会考虑到项目的规模, 也考虑到经理们的居住地离项目地点的远近. 表 E.38 给出了这些项目的预算费用 (单位: 100 万美元), 旅行时间如表 E.39 所示. 该如何指派负责这些项目的经理呢? [提示: 这里按照项目工作量所进行的基本指派定义为 (旅行时间小时 $+1$) $\times 6 \times \ln$(以 100 万美元计的项目费用)$+1$. 这个表达式常用来度量建筑项目的工作量.]

表 E.38

项目	1	2	3	4	5	6	7	8
费用 (\$$10^6$)	10	2	24	5	15	12	7	9

① 资料来源: L. LeBlanc, D. Randels, Jr., and T. K. Swann, "Heery International's Spreadsheet Optimizations Model for Assigning Managers to Construction Projects," *Interfaces*, Vol.30, No.6, pp.95-106, 2000.

表 E.39

项目	经理到各项目现场的旅行时间				
	a	*b*	*c*	*d*	*e*
1	2	5	3	1	6
2	5	4	2	5	3
3	4	1	3	2	2
4	5	3	6	3	4
5	1	4	5	6	1
6	2	4	6	2	3
7	6	7	2	3	3
8	4	2	1	5	4

第 9 章应用题

9-1 某公司每年年底都要对重型设备的状况进行检查，决定这些设备明年还能再用还是需要更换. 按规定，已经用了 3 年的设备必需要更换. 公司想要制定一项未来 10 年的设备更新计划. 表 E.40 给出了相关的数据，从第一年开始所有这些设备都是新的.

表 E.40

年度	购买价格 ($)	维修费用 ($)			折旧价值 ($)		
		0	1	2	1	2	3
1	10 000	200	500	600	9 000	7 000	5 000
2	12 000	250	600	680	11 000	9 500	8 000
3	13 000	280	550	600	12 000	11 000	10 000
4	13 500	320	650	700	12 000	11 500	11 000
5	13 800	350	590	630	12 000	11 800	11 200
6	14 200	390	620	700	12 500	12 000	11 200
7	14 800	410	600	620	13 500	12 900	11 900
8	15 200	430	670	700	14 000	13 200	12 000
9	15 500	450	700	730	15 500	14 500	13 800
10	16 000	500	710	720	15 800	15 000	14 500

第 10 章应用题

10-1 Walmark 零售商店的分销中心每天都要买进许多大宗的常用货品，因为 Walmark 下属的大量商店对这些不同的商品保持不断的需求. 过去，对订货多少和什么时候订货，完全由采购人员决定，他们的主要目的就是大量买进这些商品，以保证获得低廉的采购价格. 这种采购政策没有认真考虑过这些物品的库存的状况. 当然，关于购买量的决策考虑到了这些物品在分销中心这一层次上每年支付的总金额. 例如，若一种货品每件的采购价格是 $25，每年的消费为 10 000 件，则每年的采购支出估计达到 $250 000. 采购人员的主导思想是，一种商品每年的采购支出越高，分销中心的库存水平也应该越大. 这种指导思想就转变成了分销中心在补充订货期间所必须保持的最低库存量. 比如，对于某件商品，一个采购人员可能会每 3 个月批量采购一次.

为了更好控制库存量, Walmark 决定聘请一位运筹学专家帮助分析. 在对问题进行了研究以后, 这位专家给出了分析结论说, 分销中心大部分商品的消费率过去实际上一直是不变的, 而且 Walmark 是按照不允许出现缺货的一般性政策来运行的. 这项研究进一步说明, Walmark 所考虑的对所有商品的库存费用是单位采购价格的某个固定的百分率. 此外, 采购人员每次购买所支付的固定费用不论采购的货物如何都是一样的. 有了这些信息, 这位专家就能够对每种商品画出一条曲线, 反映出年度采购支出与订货的平均时间之间的关系, 有了这条曲线就可以决定哪些商品现在是缺货或是过剩. 这位分析人员是如何做到的呢?

10-2 某公司制造一种最终产品, 要用到某种单一零部件. 公司要从外部供应商处购买这种零部件. 市场对最终产品的需求量是一个常数, 平均每周 20 件左右. 每生产一件最终产品要用到 2 件采购来的零部件, 表 E.41 给出了库存数据.

表 E.41

	零部件	产品
每次订货的订货费 ($)	80	100
每周的单位库存费 ($)	2	5
提前时间 (周)	2	3

最终产品中未能满足市场需求的部分每周每件损失约 $8. 假定不会发生采购零部件不足的情况. 请针对零部件采购和最终产品的生产设计一项订货策略.

10-3 一家公司销售季节性商品, 每月的需求量变化很大. 表 E.42 给出了需求数据 (按件数). 由于需求的波动, 库存控制经理采取了一种季度订货方式, 订货时间为 1 月 1 日、4 月 1 日、7 月 1 日和 10 月 1 日, 按照每个季度的需求量来订货. 发出订单与到货之间的提前时间为 3 个月, 对当年需求量的估算按照年度 5 的需求量, 加上另外 10% 的安全系数.

表 E.42

	年度				
月份	1	2	3	4	5
1	10	11	10	12	11
2	50	52	60	50	55
3	8	10	9	15	10
4	99	100	105	110	120
5	120	100	110	115	110
6	100	105	103	90	100
7	130	129	125	130	130
8	70	80	75	75	78
9	50	52	55	54	51
10	120	130	140	160	180
11	210	230	250	280	300
12	40	46	42	41	43

一位新来的员工相信, 可以用基于该年平均月需求量得到的经济订货量来求出一个更好的订货策略. 需求的波动可以通过满足连续几个月的需求来进行“平滑”, 使每次的

订货量接近于经济订货量. 和经理不同, 这位新来的员工坚信, 下年度需求量的估计应根据年度 4 和年度 5 的平均值.

该公司计算库存费用时所用的每月每件库存费用为 \$0.50, 对新订单收取的订货费是 \$55.

请为该公司制定一个库存策略.

第 11 章应用题

11-1[①]某位车间经理正在对现有的一台铣床考虑 3 种方案.

(a) 用一台自动给进机 (PF) 对铣床进行改造.

(b) 购进一台带有计算机辅助设计 (CAD) 性能的新铣床.

(c) 用一套机械加工中心 (MC) 设备替代铣床.

根据两个衡量指标对这 3 个方案进行了评价: 费用和性能. 表 E.43 给出了相应的数据. 该经理设想, 费用指标要比性能指标重要 $\frac{1}{2}$ 倍. 此外, 生产速度要比准备时间重要 1 倍, 比加工废料重要 2 倍, 准备时间要比加工废料重要 3 倍. 对于费用指标, 经理估计, 维修和培训费用同样重要, 而且初始购买费用比这两项费用都重要 1 倍.

请对这个问题进行分析, 给出合理的建议.

表 E.43

评价指标	PF	CAD	MC
经费指标			
初始费用 (\$)	12 000	25 000	120 000
维修费 (\$)	2 000	4 000	15 000
培训费 (\$)	3 000	8 000	20 000
性能指标			
生产率 (件/天)	8	14	40
准备时间 (分钟)	30	20	3
废料 (磅/天)	440	165	44

11-2[②]某公司经营一种直销业务, 销售的 20 多万件商品存放在不同地区的仓库里. 过去, 公司认为, 掌握每个仓库库存量的准确记录非常重要. 因此, 每年都下令对全部的库存进行盘点, 这是一项繁重而且大家不愿意做的工作. 所有仓库都不得不勉强完成. 该公司在每次盘点后, 对每个仓库随机抽检 100 件产品, 检查每个地区的物流运作情况. 检查结果表明, 平均来讲, 每个仓库仅有 64% 的货品符合实际库存情况, 这种状况简直不可接受. 为了采取补救措施, 公司下令对高价的和流动性强的货品进行经常性的库存盘点. 公司聘请了一位系统分析人员负责制定对这些货品进行跟踪检查的制度.

① 资料来源: S. Weber, "A Modified Analytic Hierarchy Process for Automated Manufacturing Decisions," *Interfaces*, Vol.23, No.4, pp.75-84, 1993.

② 资料来源: I. Millet, "A Novena to Saint Anthony, or How to Find Inventory by Not Looking," *Interfaces*, Vol.24, No.2, pp.69-75, 1994.

这位系统分析人员没有直接按照公司的要求找出要跟踪的货品，而是决定先找出问题的根源. 根据发现的问题，他决定把研究的目标从"如何提高库存盘点的频度?"变成"如何提高库存盘点的准确度?". 这项研究有了如下分析结果：假定每个仓库中准确盘点的货品比例为 p，那么就有理由相信，有 95% 的机会使得第一次盘点时是正确的某件货品在以后的再盘点中仍然是准确的. 对于第一轮盘点不准确的比例 $1-p$，再次盘点正确的机会为 80%. 利用这一信息，这位分析人员建立了一个决策树，用来画出一个损益平衡图，对第一轮和第二轮盘点的准确度进行比较. 最终得出的结果是，那些准确度水平大于该损益平衡阈值的仓库不必要再对库存进行盘点. 对所提出解决方案的这一出乎意料的结果致使每个仓库都积极努力，尽量在第一轮盘点中得出正确的结果，并且所有仓库都大大提高了盘点准确度.

这位分析人员是如何向管理部门说明他提出的再盘点控制阈值是管用的呢?

11-3[①]在航空业界，飞行人员的持续工作小时数受工会协议的限制. 特别地，波音 747 航班的最长飞行小时不应超过 16 个小时，波音 707 航班的最长飞行时间应在 14 小时之内. 假如由于意外延误造成超出这些限制的话，就要有新的机组来替换. 对于可能出现的这种情况，一般航空公司都保留了备份机组执行飞行任务. 一个备份机组人员每年的费用大约为 3 万美元. 反过来，由于没有备份机组造成的隔夜延误航班的损失一次可达 5 万美元. 一位机组人员每周可值班 4 天，每天 12 小时，其他 3 天不能执班，必须休息. 一架 B-747 航班也可以由两名 B-707 飞行员执行飞行任务.

表 E.44 是根据过去 3 年的历史数据统计的备份机组的值班概率. 表上数据说明，对于一个 14 小时的航班，B-747 飞行员的值班概率是 0.014，而 B-707 备份机组的值班概率为 0.072. 表 E.45 给出了一个典型的高峰日的航班时刻. 针对备份机组的现行政策要求，在 5:00 到 11:00 之间需要利用 2 个 (7 人) 备份机组，11:00 到 17:00 之间要有 4 个备份机组，17:00 到 23:00 间要用 2 个备份机组.

请评估现行备份机组政策的有效性. 特别是，分析现行要求的备份机组数是太多、太少、还是刚好合适?

表 E.44

航班类别	飞行小时	值班概率	
		B-747	B-707
1	14.0	0.014	0.072
2	13.0	0	0.019
3	12.5	0	0.006
4	12.0	0.016	0.006
5	11.5	0.003	0.003
6	11.0	0.002	0.003

① 资料来源：A. Gaballa, "Planning Callout Reserves for Aircraft Delays," *Interfaces*, Vol.9, No.2, Part 2, pp.78-86, 1979.

表 E.45

时间	飞机型号	航班类别
8:00	707	3
9:00	707	6
	707	2
10:00	707	3
11:00	707	2
	707	4
15:00	747	6
16:00	747	4
19:00	747	1

11-4[①]在轰动一时的 1982 年 John Hinkley 刺杀美国总统里根案件期间, 辩护律师想要借用 Hinkley 的 CAT 扫描结果来证明他的当事人精神有问题. Hinkley 的 CAT 扫描片子并没有表明他的大脑有萎缩, 庭审期间的专家证词认为, 在被诊断患精神分裂症的患者中, 有 30% 的人出现大脑萎缩; 而在非精神分裂症患者中只有 2% 的人会出现大脑萎缩症. 统计结果显示, 大约有 1.5% 的美国人患有精神分裂症.

从提供 CAT 扫描结果作为证据对案件审理结论影响程度的角度, 对这一问题进行分析.

11-5[②]一位教授要估算他班上的大三和大四高年级学生在大学期间的考试中曾经作弊的概率. 为了从学生那里得出无偏的 (真实的) 的答案, 他要求每个学生自己投掷一枚硬币, 如果正面朝上, 要回答一个圈套问题; 如果是正面朝下, 就要回答一个真实问题. 真实问题是"你曾经在考试中作过弊吗?", 而圈套问题为"你是即将毕业的大四学生吗?"每个学生在一张纸上回答"是"或"否", 然后收回这张纸, 由教授来统计. 答案是可以保密的, 因为只有学生自己知道他回答了哪一个问题. 在 35 名参与了这项试验的学生中, 有 20 名大四学生. 试验统计结果表明, 有 18 个回答"是", 17 个回答"否". 利用这些信息估计一名学生在该班的考试中曾经作弊的概率.

第 12 章应用题

12-1[③]Elkins 银行现开办一种传统的不下车银行服务站和两个"气动"车道, 通过一个气送管道连接到银行内部. 该银行想要扩充现有设施, 以便来到的车辆可以在平均 4 分钟内办理完一笔业务. 这一时间限制的设定是根据对心理学的研究, 研究表明, 顾客的忍耐程度通常不超过分针移动两格, 在大多数手表上代表 5 分钟. 为了收集必要的数据, 银行研究小组观察了现有柜台的运行情况. 研究一段时间后, 有位小组人员注意到, 一位顾客花在进入

① 资料来源: A. Barnett, I. Greeberg, and R. Machol, "Hinckley and the Chemical Bath, "*Interfaces*, Vol.14, No.4, pp.48-52, 1984.

② 资料来源: R. Sheaffer, J. Witmer, A. Watkins, and M. Gnanadesikan, *Activity-Based Statistics*, Springer-Verlag, New York, 1996, p.133.

③ 资料来源: B. Foote, "A Queuing Case Study in Drive-In Banking," *Interfaces*, Vol.6, No.4, pp.31-37, 1976.

车道的时间与柜员完成一笔业务所花的时间有着明显的差异. 事实上, 一辆车花在该系统中的时间构成为 (1) 把车辆开到队列, (2) 移动到柜台窗口, (3) 告诉柜员服务要求, (4) 柜员根据要求进行服务, (5) 把车辆开走. 在这个时间段的第一、第二和第五部分, 柜员没有什么事做. 实际上, 柜台服务人员只有 40% 的时间忙于服务顾客. 根据这些信息, 研究小组发现了现有系统尚有减少运行成本的余地.

为了改进现有汽车银行系统的运行, 研究小组的建议会是什么呢? 讨论这一建议的所有含义.

12-2 一家国营反虐待儿童救助中心每天的办公时间从上午 9 点开门, 到晚上 9 点结束. 报告虐待儿童案件的电话完全随机打来, 表 E.46 给出了 7 天期间每小时打进的报案数. 该表并不包括由于电话占线所没有打进的电话, 每个电话报案的时间长度也是随机的, 最长 12 分钟, 平均 7 分钟. 过去的记录表明, 该中心电话报案数每年增长 15%.

该中心希望确定必需安装的电话线路数, 以为现在和将来提供足够的服务. 尤其是要注意到, 尽量减少由于线路占线所带来的负面影响.

表 E.46

开始时间	每天电话报案数						
	1	2	3	4	5	6	7
9:00	4	6	8	4	5	3	4
10:00	6	5	5	3	6	4	7
11:00	3	9	6	8	4	7	5
12:00	8	11	10	5	15	12	9
13:00	10	9	8	7	10	16	6
14:00	8	6	10	12	12	11	10
15:00	10	9	12	4	10	6	8
16:00	8	6	9	14	12	10	7
17:00	5	10	10	8	10	10	9
18:00	5	4	6	5	6	7	5
19:00	3	4	6	2	3	4	5
20:00	4	3	6	2	2	3	4
21:00	1	2	2	3	3	5	3

12-3 某制造公司用 3 辆卡车在 6 个车间之间运送原材料. 卡车的用户一直要求车队增加第 4 辆卡车, 以缓解送货延误过长的问题. 这些卡车并没有一个总站, 管理部门认为, 让卡车在工厂不断移动可能效率更高. 要求用卡车的车间必需等待附近的车辆到达. 在有车的情况下, 就会响应请求; 反之, 车间必需等待另一辆卡车出现. 表 E.47 给出了每个小时的用车请求频率. 对每个车间的服务时间 (按分钟) 基本相同. 表 E.48 为一个典型服务时间的分布情况.

分析现行运作方式的有效性.

表 E.47

请求/小时	频率	请求/小时	频率
0	30	7	47
1	90	8	30
2	99	9	20
3	102	10	12
4	120	11	10
5	100	12	4
6	60		

表 E.48

服务时间 t	频率	服务时间 t	频率
$0 \leqslant t < 10$	61	$50 \leqslant t < 60$	4
$10 \leqslant t < 20$	34	$60 \leqslant t < 70$	4
$20 \leqslant t < 30$	15	$70 \leqslant t < 80$	3
$30 \leqslant t < 40$	5	$80 \leqslant t < 90$	2
$40 \leqslant t < 50$	8	$90 \leqslant t < 100$	2

12-4 Jon Micks 是一位年轻的工业工程师, 最近应聘到 Metalco 公司任职. 该公司拥有一个 30 台机器的工厂, 雇用了 6 名修理工负责机器维修. 工厂上一天班, 从早上 8 点开始, 到下午 4 点结束. Jon 的第一项任务是分析一下工厂维修服务的有效性. 为了完成这项任务, 他收集了 3 台随机抽取的机器维修日志数据, 如表 E.49 所示. 另外, 还检查了 5 天随机选择日期的维修记录, 编辑成了表 E.50 的数据, 代表工作日每小时开始时的故障机器数 (包括正在修理的机器).

表 E.49

5 号机器		18 号机器		23 号机器	
故障时间	完工时间	故障时间	完工时间	故障时间	完工时间
8:05	8:15	8:01	8:09	8:45	8:58
10:02	10:14	9:10	9:18	9:55	10:06
10:59	11:09	11:03	11:16	10:58	11:08
12:22	12:35	12:58	13:06	12:21	12:32
14:12	14:22	13:49	13:58	12:59	13:07
15:09	15:21	14:30	14:43	14:32	14:43
15:33	15:42	14:57	15:09	15:09	15:17
15:48	15:59	15:32	15:42	15:50	16:00

表 E.50

日期	每小时故障机器总数							
	8:00	9:00	10:00	11:00	12:00	13:00	14:00	15:00
10/2	6	6	9	6	8	8	7	7
10/29	9	8	5	9	5	5	6	8
11/4	6	6	5	7	7	8	6	5
12/1	9	5	9	7	5	7	5	5
1/19	6	5	8	5	9	8	8	6

Jon 就去见了上司 Becky Steele, 汇报他收集的数据情况. 声称他相信, 该车间的机器故障/修理过程是完全随机的, 这个问题可以用泊松队列来描述. 基于在该车间多年的工作经验, Becky 也认为该问题完全是随机的. 基于这种情况, Becky 仔细观察了 Jon 的数据, 经过计算后 Becky 声称, 数据是有问题的. Becky 是如何得出这一结论的呢?

12-5 Yellow Cab 出租车公司拥有 4 辆出租车, 每天提供 10 个小时的租车服务. 调度站收到的叫车请求服从泊松分布, 平均每小时 20 个请求. 乘车时间长度已知为指数分布, 平均值为 11.5 分钟. 由于叫车的人比较多, 公司限制调度站的等待队列最多 16 位顾客. 一旦达到上限, 就建议后面的顾客到别的地方去叫车, 避免等待时间过长.

该公司经理 Kyle Yellowstone 恐怕这样会丧失太多的业务, 所以他考虑扩大车队规模. Yellowstone 估计从每位乘客挣到的收入平均为 \$5, 一辆新车的购置费用是 \$18 000. 一辆新车按 5 年使用期, 然后还可以按 \$3 500 卖出. 保养和运行一辆出租车的费用是 \$20 000. Yellowstone 先生增加车队规模合算吗? 如果可行, 要增加多少辆车? 为分析方便, 假定每年的利率为 10%.

第 13 章的综合问题

13-1 假如给定 3 个点:

$$A = (6, 4, 6, -2), \quad B = (4, 12, -4, 8), \quad C = (-4, 0, 8, 4)$$

建立一个系统化的方法, 能够判断下列每个点是否可表示为 A, B, C 的一个凸组合:

(a) $(3, 5, 4, 2)$ **(b)** $(5, 8, 4, 9)$

13-2 考虑线性规划问题:

$$\begin{aligned}
\max \quad & z = 3x_1 + 2x_2 \\
\text{s.t.} \quad & x_1 + 2x_2 \leqslant 6 \\
& 2x_1 + x_2 \leqslant 8 \\
& -x_1 + x_2 \leqslant 1 \\
& x_1, \quad x_2 \geqslant 0
\end{aligned}$$

做出最优解单纯形表 (为方便起见, 可以用 TORA 程序), 然后直接用最优解单纯形表的信息求出该问题的次最优的极点解 (相对于"绝对"最优值), 并用图解法对该问题的答案进行验证 (提示: 考虑最优解附近的极值点.)

13-3[①] 例 2.3-9 的修剪损失模型中假定, *所有的*切刀设置都是事先确定好的. 在实际中, 可以用带有嵌入整数线性规划的修正单纯形法, 并利用下面的列生成方法来产生希望的切刀设置.

---

① 资料来源: P. Gilmore and R. Gomory, "A Linear Programming Approach to the Cutting Stock Problem," *Operations Research*, Vol.9, No.4, pp.849-859, 1961. —, "A Linear Programming Approach to the Cutting Stock Problem:Part II," *Operations Research*, Vol.22, No.4, pp.863-888, 1963.

第 0 步 (开始求解) 选择显然的初始基本可行解, 每个所需要的宽度恰好由一卷纸组成. 例如, 在例 2.3-9 中, 基本解 $\boldsymbol{X}_{\boldsymbol{B}_0} = (x_1, x_2, x_3)^{\mathrm{T}}$ 对应于下面的基:

$$\boldsymbol{B}_0 = \begin{matrix} x_1 & x_2 & x_3 & \\ \left(\begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix}\right) & & & \begin{matrix} \leftarrow 5\text{ ft} \\ \leftarrow 7\text{ ft} \\ \leftarrow 9\text{ ft} \end{matrix} \end{matrix}$$

这个基对应于切出 150, 200, 300 个标准纸卷, 以分别得到 5 英尺、7 英尺和 9 英尺的纸卷. 在这一点上, 有多少料头并不重要. 利用修正单纯形法的符号, 这个解相当于:

$$\boldsymbol{B}_0 = \boldsymbol{B}_0^{-1} = \boldsymbol{I}$$

$$\boldsymbol{c_B B_0^{-1}} = (1, 1, 1)$$

第 j 步 (希望的切刀设置): 令 $\boldsymbol{B}_{j-1}^{-1}$ 为关于切刀设置组合 $j-1$ 的基矩阵的逆. 第 j 个切刀设置可用列向量 $\boldsymbol{P}_j = (a_1, a_2, \cdots, a_m)^{\mathrm{T}}$ 来表示, 其中 m 是能从一个标准纸卷切出特殊纸卷的数目, a_i 是纸卷类型 i 的数目, $i = 1, 2, \cdots, m$. 例如在例 2.3-9 中, $m = 3$, 设置 $(2, 1, 0)$ 得到 2 个 5 英尺的纸卷, 1 个 7 英尺的纸卷, 以及 0 个 9 英尺的纸卷. 我们的目标是求出 $\boldsymbol{P}_j$ 的元素, 使得相应的变量 x_j 在下一次迭代中为基变量. 这就等价于, 在已知相应的切刀设置 $a_1, a_2, \cdots, a_n$ 可行的情况下, 求得到最大简约费用 $z_j - c_j$ 的 $\boldsymbol{P}_j$(请注意, 修剪损失模型是一个最小化问题). 令 W 为标准纸卷的宽度, $\boldsymbol{w} = (w_1, w_2, \cdots, w_m)$ 表示有序纸卷的宽度. 因此, 确定最有希望的 $\boldsymbol{P}_j$ 就相当于求解下面的整数线性规划问题:

$$\begin{aligned} \max\ & z_j - c_j = \boldsymbol{c_B B}_{j-1}^{-1} \boldsymbol{P}_j - c_j \\ \text{s.t.}\ & \boldsymbol{w P}_j \leqslant W \\ & \boldsymbol{P}_j \geqslant \boldsymbol{0}\ \ \text{为整数} \end{aligned}$$

修剪损失问题 $c_j = 1, \forall\, j$, 因此 $\boldsymbol{c_B} = (1, 1, \cdots, 1)$. 令 $(d_1, d_2, \cdots, d_m) = \boldsymbol{c_B B}_{j-1}^{-1}$ 表示相应于基 $\boldsymbol{B}_{j-1}^{-1}$ 的对偶变量, 可以写出下列 ILP 问题:

$$\begin{aligned} \max\ & z = \sum_{k=1}^{n} d_k a_k \\ \text{s.t.}\ & \sum_{k=1}^{n} w_k a_k \leqslant W \\ & \sum_{k=1}^{n} d_k a_k \geqslant 1 + \epsilon \\ & a_k \geqslant 0\ \ \text{为整数} \end{aligned}$$

第 2 个约束要求, 为了让 $\boldsymbol{P}_j$ 为一个有希望的进基向量, 变量 $z_j \left(= \sum_{k=1}^{m} d_k a_k\right)$ 必须严格大于$c_j (= 1)$ (ϵ 是一个充分小量).

从这个 ILP 的解可以得出两个结论:

(1) 如果解是可行的, 则 P_j 一定是基. 用修正单纯形法求离基向量, 得到新的逆矩阵 B_j^{-1}. 重复第 j 步.

(2) 若问题没有可行解, 则不存在其他有希望的 P_j, 且上一个基本解就是最优的.

用上述给定的算法求解例 2.3-9 的修剪损失问题. 为了方便, 文件 amplPob7-3.txt 给出了一个求解 ILP 的 AMPL 模型. 虽然现在还没有讲到 ILP 算法 (见第 8 章), 但这个 AMPL 模型基本上还是一个线性规划, 只是另外要求变量 a_i 取整数值, 即

```
var a(1..3)>=0, integer;
```

此外, 语句

```
option solver cplex;
```

必须放在命令

```
solve;
```

之前.

两步迭代之间所改变的数据只有对偶变量值. 可以用 AMPL 的 `let` 命令交互式地改变这些数值. 例如, 假如我们要把 d_1 和 d_2 的值分别改成 0.5 和 0.75(其他值保持不变)的话, 则可以交互式地输入

```
ampl: let d[1]=.5;
ampl: let d[3]=.75;
ampl: solve; dispay a;
```

这意味着不需要每次迭代都要把这些数据编写在模型里.

13-4 区间规划. 考虑下面的线性规划问题:

$$\max\ z = \{\boldsymbol{CX} | \boldsymbol{L} \leqslant \boldsymbol{AX} \leqslant \boldsymbol{U}, \boldsymbol{X} \geqslant \boldsymbol{0}\}$$

其中 $\boldsymbol{L}$ 和 $\boldsymbol{U}$ 为常数列向量. 定义松弛向量, 使得 $\boldsymbol{AX}+\boldsymbol{Y}=\boldsymbol{U}$. 证明这个线性规划等价于

$$\max\ z = \{\boldsymbol{CX} | \boldsymbol{AX}+\boldsymbol{Y}=\boldsymbol{U}, 0 \leqslant \boldsymbol{Y} \leqslant \boldsymbol{U}-\boldsymbol{L}, \boldsymbol{X} \geqslant \boldsymbol{0}\}$$

用上述办法求下列线性规划:

$$\begin{aligned}
\min\quad & z = 5x_1 - 4x_2 + 6x_3 \\
\text{s.t.}\quad & 20 \leqslant x_1 + 7x_2 + 3x_3 \leqslant 46 \\
& 10 \leqslant 3x_1 - x_2 + x_3 \leqslant 20 \\
& 18 \leqslant 2x_1 + 3x_2 - x_3 \leqslant 35 \\
& x_1,\ x_2,\ x_3 \geqslant 0
\end{aligned}$$

13-5 给定第 13-2 题的最优解为 $x_1=\frac{10}{3}$, $x_2=\frac{4}{3}$, $z=\frac{38}{3}$. 若给出 $x_1=\frac{10}{3}+\theta$, θ 无符号限制, 请画出最优解 z 随 θ 变化的曲线. 注意到 $x_1=\frac{10}{3}+\theta$ 随 x_1 的轨迹在最优值上下移动.

13-6 考虑下面的求最小值的线性规划：

$$\begin{aligned} \min\ z &= (10t-4)x_1+(4t-8)x_2 \\ \text{s.t.}\ & 2x_1+2x_2+x_3 = 8 \\ & 4x_1+2x_2+x_4 = 6-2t \\ & x_1,\ x_2,\ x_3,\ x_4 \geqslant 0 \end{aligned}$$

其中 $-\infty < t < \infty$. 对这个问题的参数分析得出以下结果：

$$\begin{aligned} -\infty < t \leqslant -5: &\quad \text{最优基为 } \boldsymbol{B}=(\boldsymbol{P}_1,\boldsymbol{P}_4) \\ -5 \leqslant t \leqslant -1: &\quad \text{最优基为 } \boldsymbol{B}=(\boldsymbol{P}_1,\boldsymbol{P}_2) \\ -1 \leqslant t \leqslant 2: &\quad \text{最优基为 } \boldsymbol{B}=(\boldsymbol{P}_2,\boldsymbol{P}_3) \end{aligned}$$

求出对于 $t \geqslant 2$ 而言可能存在的 t 的所有临界值.

13-7 假设最优值线性规划表示为

$$\begin{aligned} \max\ z &= c_0 - \sum_{j\in NB}(z_j-c_j)x_j \\ \text{s.t.}\ x_i &= x_i^* - \sum_{j\in NB} a_{ij}x_j,\quad i=1,2,\cdots,m \\ & \forall\ x_i,\ x_j \geqslant 0 \end{aligned}$$

这里 NB 是非基变量的集合. 假设, 我们对于现行基变量 $x_i = x_i$ 设定一个限制条件 $x_i \geqslant d_i$, 其中 d_i 是大于 x_i 的最小整数. 当把这一约束加在问题里后, 估计 z 的最优值上界. 假定设置的限制条件是 $x_i \leqslant e_i$, 其中 e_i 是小于 x_i 的最大整数, 重复同样的过程.

第 15 章应用题

15-1[①]一家电话公司经营若干个电话中心, 为各分片顾客提供上门装机服务. 共有 60 种型号的电话机供顾客选择, 现在, 每个电话中心保存 15 到 75 天不等的话机库存. 管理部门认为, 这样的库存水平太多, 因为他们每天都从中心仓库补充. 同时, 管理部门想要保持每个电话中心的充足库存量, 以满足 95% 的顾客要求. 一个研究小组从收集相关数据开始对这个问题进行研究. 该小组的目的是对每个电话机型号建立一个最优的库存水平. 表 E.51 表示一天内安装的绿色、桌面和拨号盘 (绿色 500) 型电话机的数量. 对所有的电话机型号都编制了这样的表格.

表 E.51

安装型号	0	1	2	3	4
装机量	189	89	20	4	1

用于确定每种电话型号最优库存水平的费用参数很难估计, 因此难以使用传统的库存模型. 根据观察发现, 回归和时间序列分析都不能对需求趋势给出满意的预测, 该研究小组决定利用一种更基本的方法对不同型号的电话计算出适当的库存水平.

① 资料来源：R. Cohen, and F. Dunford, “Forecasting for Inventory Control: An Example of When ‘Simple’ Means ‘Better’,” *Interfaces*, Vol.16, No.6, pp.95-99, 1986.

请为该小组建议一种方法，用于求出不同型号电话足够的库存量，并给出决策所需要的所有假设条件.

15-2[①]一家小商店的库存经理订货时，只考虑价格优惠或利用一家供应商提供多个订单机会. 结果发现，订货量和订货周期 (连续两次订货之间的间隔) 都变成随机的了. 此外，由于该经理的订货策略大部分受到非库存因素的影响，订货量和周期长度可以认为是无关的，即较短的周期长度并不一定订货量少，反之亦然.

表 E.52 给出了同批订货的 3 种货品的典型数据. 这些数据表明，订货量和周期长度都是随机的. 此外，粗略查看表中的数据可发现，订货量和周期长度之间缺乏相关性.

表 E.52

周期长度 (月)	订货量 (件)			周期长度 (月)	订货量 (件)		
	1	2	3		1	2	3
2.3	10	8	1	1.6	2	0	0
2.6	4	6	0	0.5	5	3	1
4	1	4	2	2.1	10	7	2
2.0	8	6	2	2.3	4	12	4
1.2	7	0	2	2.4	8	9	3
1.4	0	10	1	2.1	10	8	5
1.7	1	2	0	2.2	9	13	2
1.3	0	5	2	1.8	12	8	4
1.1	9	4	3	0.7	6	4	2
1.8	4	6	2	2.1	5	4	0

对整个这批数据进行拟合分析发现 (见第 14 章)，这 3 种货品的需求率 (订货量除以周期长度) 服从韦布尔分布，密度公式 $f(r)$ 为

$$f(r)=\frac{2r}{\alpha}\mathrm{e}^{-r^2/\alpha}, r\geqslant 0$$

其中 r 是该货品的需求率. 类似地，分析表明，周期长度的倒数 $s(x)$ 的分布为指数形式

$$s(x)=\beta\mathrm{e}^{-\beta(x-a)}, x\geqslant a$$

这里 a 假定为 x 的最小值.

最优订货量可以根据求每月的期望最大利润来确定，期望利润定义如下：

$$\begin{aligned}\text{期望利润}&=\int\left\{\frac{1}{t}\int u(q,r,t)f(r)\mathrm{d}r\right\}g(t)\mathrm{d}t\\&=\int\left\{x\int u\left(q,r,\frac{1}{x}\right)f(r)\mathrm{d}r\right\}s(x)\mathrm{d}x\end{aligned}$$

其中 t 和 $g(t)$ 为周期长度及其密度函数. 利润函数 $u(q,r,t)$ 根据货品的单位净利润 p、每件货品每月存货费用 h，以及固定订货费 K 计算.

① 资料来源：A. Holt, "Multi-Item Inventory Control for Fluctuating Reorder Intervals," *Interfaces*, Vol.16, No.3, pp.60-67, 1986.

(a) 利用这 3 种货品的数据, 求每个需求率的概率密度函数.

(b) 用周期长度数据求 $s(x)$.

(c) 建立 $u(q,r,t)$ 的数学表达式.

根据费用数据 $p_1=\$100$, $p_2=\$150$, $p_3=\$125$, $h_1=\$2$, $h_2=\$1.2$, $h_3=\$1.65$, $K=\$30$, 求出这 3 种货品的最优订货量.

第 20 章应用题

20-1[①]某大学工业工程系有 3 位教授, 为两个学期的学年提供 5 门课程的教学. 该系有 2 名研究生, 可教课程 C1 和 C3, 但只有当教授不能讲这些课程时才轮到他们上课. 每个研究生每学期最多能教一门课. 表 E.53 和表 E.54 指定了每位教授愿意讲授的课程和每门课程每学年的教学量.

建立一个模型, 用于将教授 (必要时包括研究生) 分配到指定的课程.

表 E.53

课程	每学年教学量	每学期教学量	
		秋季	春季
C1	2	1	1
C2	2	1 或 2	1
C3	2	1 或 2	1 或 2
C4	1	1	0
C5	1	0	1

表 E.54

教授	每学年教学量	每学期教学量		对课程的喜好顺序
		秋季	春季	
P1	1	0 或 1	0 或 1	C1>C2>C5
P2	3	1 或 2	1 或 2	C1>C3>C2>C4
P3	2	0,1,2	0 或 1	C5>C4>C3>C1

第 21 章应用题

21-1 一项公开辩论称, 美国高中学生的学习能力倾向测验 (SAT) 近来的平均分数上升是由于统计原因造成的, 而不反映教学方法的改进. 尤其是认为, 每个家庭子女数的减少创造了一种环境, 让孩子们有更多的时间跟成年人 (即父母) 在一起, 从而提高了他们的智力水平. 相反, 大家庭的孩子就没有这种智力上的"优势", 因为会受到他们兄弟姐妹不成熟的影响.

对于这种争议, 如何建立对 SAT 成绩的预测回归方程?

第 22 章应用题

22-1 UPPS 公司用卡车向顾客运送订单货物. 该公司想要指定一项未来 5 年的车辆更新策略. 一辆新卡车的每年运行费用服从正态分布, 平均值为 \$300, 标准差为 \$50. 运行费的平

① 资料来源: J. Dyer and J. Mulvey, "An Integrated Infomation/Optimization for Academic Planning," *Management Science*, Vol.22, No.12, pp.582-600, 1976.

均值和标准差以后每年增加 10%. 一辆新车的现行价格是 \$20 000, 预计每年涨价 12%. 由于卡车的使用强度很高, 有可能随时报废. 一辆卡车的回收价值要根据它是否有故障而定. 从第 6 年开始卡车就要报废, 报废的价值仍然取决于车况 (是否还能开). 表 E.55 给出了与卡车使用年限相关的数据.

表 E.55

卡车年限	0	1	2	3	4	5	6
故障概率	0.01	0.05	0.10	0.16	0.25	0.40	0.60

如果这辆卡车仍然能开, 运行 1 年的二手价值是新车购置价格的 70%, 其后每年降低 15%. 如果车有故障不能开了, 则折价一半. 从第 6 年开始的报废回收价, 如果还能开, 则报废价为 \$200, 若是坏车, 则为 \$50, 请为该卡车建立一个最优更新策略.

索　　引